AD_____

Award Number DAMD17-98-1-8039

TITLE: FACTS (Find the Appropriate Clinical TrialS) for You: A Computer-Based Decision Support System for Breast Cancer Patients

PRINCIPAL INVESTIGATOR: Lucila Ohno-Machado, M.D.,Ph.D.

CONTRACTING ORGANIZATION: Brigham and Women's Hospital
Boston, Massachusetts 02115

REPORT DATE: May 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1999	Annual (20 Apr 98 - 19 Apr 99)

4. TITLE AND SUBTITLE
FACTS (Find the Appropriate Clinical TrialS) for You: A Computer-Based Decision Support System for Breast Cancer Patients

5. FUNDING NUMBERS
DAMD17-98-1-8039

6. AUTHOR(S)
Lucila Ohno-Machado, M.D.,Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Brigham and Women's Hospital
Boston, Massachusetts 02115

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We have developed a system for clinical trial eligibility determination where patients or primary care providers can enter clinical information about a patient and obtain a ranked list of clinical trials for which the patient is likely to be eligible. We used clinical trial eligibility information from the National Cancer Institute's Physician Data Query (PDQ) database. We translated each free-text eligibility criterion into a machine executable statement using a derivation of the Arden Syntax. Clinical trial protocols were then structured as collections of these eligibility criteria using XML. The application compares the entered patient information against each of the eligibility criteria and returns a numerical score. Results are displayed in order of likelihood of match. We have tested our system using all phase II and III clinical trials for treatment of metastatic breast cancer found in the PDQ database. Preliminary results are encouraging.

We have completed the tasks in the approved statement of work that were planned for Year 1 of this project. A working prototype of the system we envision as our final product is available at http://telmato.bwh.harvard.edu:8000/FACTS/FormIn.htm.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
82

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

__✓__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Luis Otavio Machado 5/15/99

PI - Signature Date

Table of Contents

Introduction

Although participation in clinical trials has been shown to improve health outcomes, accrual of patients is difficult and is estimated to be below 5% of the eligible population. Lack of information and automated tools to search clinical trials appropriate for each particular patient are some of the main reasons for low accrual. The purpose of this project is to build and evaluate a computer-based decision support system to help patients and primary care providers seek appropriate trials for their specific situation, even in conditions of uncertainty (missing data). We have proposed to make available, via the WWW, a search engine for clinical trial eligibility that searchers trials listed in the PDQ database of the NCI. On-line description of the project and working prototype can be found from http://dsg.harvard.edu/public/dsg/projects/facts.html

Body

Briefly, we have proposed to build our computer-based eligibility determination engine in two stages: (1) build an ad-hoc deterministic (i.e., nonprobabilistic engine not able to deal with uncertainty or consider associations among eligibility criteria and patient data values), and (2) build a probabilistic engine, based on belief networks, that is able to statistically imput values for missing data, given the information it can gather from the patient or health care provider, and can take into account associations among variables and patient data values.

An overall summary of the goals and accomplishments of this project until February, 1999, is given in [Ohno-Machado et al., 1999 in Appendix 1]. Additional work has been developed since then, and is included in the summary below, in which a description of the research accomplishments associated with each Task outlined in the Approved Statement of Work (restated in **bold** face) are given:

Task 1. **Analyze, structure, and construct data entry forms for eligibility criteria derived from clinical trials for breast cancer treatment available in PDQ, Months 1-6:**

a. **PDQ clinical trial summaries for health care professionals will be dissected**

We have analyzed and encoded 85 clinical trials from the PDQ database. These are all Phase II and Phase III trials for the treatment of metastatic or recurrent breast cancer. A total of 2188 criteria in these trials were encoded. The medium number of crtieria per protocl was 25, with numbers ranging from 6 to 45. The encoded criteria were structured in the format described in the next item. The encoded protocols can be found in: http://telmato.bwh.harvard.edu:8000/FACTS/data/

An example of an encoded protocol is given in the Appendix 6. The variables utilized in coding all protocols are shown in the Appendix 7.

b. A structured format for storing eligibility criteria in a relational database will be defined

We have evolved three different structure formats for storing eligibility criteria into databases:

(1) A flat file-based collection of discrete, non-overlapping fields for use in simple spreadsheets

(2) An ad-hoc syntax suitable for representation inside Access queries

(3) An extension of the standard Arden syntax for representing criteria in medical logic modules

The structures in (1) and (2) were created to perform simple queries on the data, for a first pass on counting and characterizing which criteria were most prevalent. These structures were simple to handle, but proved to be insufficient for handling certain types of criteria, such as those requiring a hierarchical structure, those that could be inferred from other criteria, and those involving temporal sequences.

Our extension to Arden, briefly documented in [Wang et al, 1999 in Appendix 2], was created to circumvent these problems. We used the eXtensible Markup Language (XML) to represent the structure of the protocols in terms of objects and attributes. The files were stored first in a Linux server, then moved to a Windows NT server. The Document Type Definition for the protocol files and the variables utilized in them are shown in the Appendices 4 and 5.

c. WWW-based data entry forms will be constructed an linked to database

We also developed three versions of these WWW-based forms.

(1) Static forms featuring the most frequent of easy to collect criteria. Implemented using HTML forms connected by CGI scripts to a relational database residing on a Linux server.

(2) Static forms featuring the most frequent criteria plus dynamic forms featuring the criteria that would likely be most informative for a particular patient. Implemented as Active Server Pages linked to a database of XML files representing the protocols.

(3) Graphical User Interface (GUI) was modified to be more visually appealing.

d. Database for interim storage of patient data will be constructed

A simple structured format was used to store patient data and communicate it to the WWW-server. We used XML files for this purpose. Since we plan to link to a real patient database in the future, and this database is supposed to be interim storage, lasting one WWW session only, detailed elaboration of structure for the patient data was not necessary.

Task 2. **Construct simple models that do not model uncertainty to assess the need for belief network models, Months 7-9:**

a. **Simple rule-based system construction using knowledge from domain expert**

The rule-based system used for this phase consisted of the following. Protocols for which the patient had one or more values that met exclusion criteria (or did not meet inclusion criteria) were removed from consideration for that patient. For the remaining protocols, the ranking took into account how many criteria still needed to be met. A protocol with several criteria to be met would rank lower than one with just one or two criteria. The importance of each criterion was not taken into account, nor the dependencies among criteria.

b. Preliminary evaluation of simple rule-based system

We have consulted with our two oncologists to assess how this simple system was performing, both in terms of adequacy of trials selected and usability of the interface. The experts provided useful suggestions, which were incorporated. We have also presented the system to a few colleagues, obtaining positive feed-back.

Task 3. **If results from Task 2 show that belief networks are needed, construct belief network to model uncertainty in most common eligibility criteria and perform inference on entered data, else refinement of simple models and interface construction will take place, Months 9-12:**

a. Belief network model will be constructed using knowledge from domain expert

A simple Belief network featuring relations among laboratory values that are frequently encountered in eligibility criteria was constructed.

b. Belief network model will be integrated with WWW and database environments to create application

We have investigated the ways by which the Belief network software acquired for this project, Netica, from Norsys Inc., can be adapted to work with the current eligibility engine engine. The simple model is not detailed enough to warrant immediate integration with the existing, functional prototype. We have decided to wait until the current work on a more detailed model is finished to proceed with this step.

c. Algorithm for ranking possible trials for a patient will be implemented

This ranking algorithm was implemented, though not considered completely satisfactory because it does not take into account dependencies among variables, nor importance or "mutability" of values matched against eligibility criteria. For example,

d. GUI for displaying results and linking to specific summaries in PDQ will be built

This graphical user interface was built using HTML, and later refined with ASP. It summarizes the data entered for a given patient, displays protocols in reverse order of

appropriateness (i.e., most appropriate trials are listed first), and suggests which variables should be asked next so that the maximum number of remaining trials can be triaged out of the list. The GUI is usable by health care professionals, but still uses medical terminology for several criteria (e.g., refers to ALT levels instead of easier to understand proxies, such as "liver function"). We are consulting with patient advocacy groups to assess the usability of the current interface and determine the best proxies for certain criteria if they are not intelligible for the general public.

We are currently in the process of changing the graphical appearance of the interface.

Task 4. **Redesign of evaluation methods and interim analysis and system refinement, Months 12-24:**

a. Evaluation methodology will be redesigned

Redesign of the evaluation strategy has started given feedback from oncologists from the Dana-Farber Cancer Institute and consultants from the biostatistics service at Brigham and Women's hospital (Dr. Robert Lew, PhD). Retrospective data from Brigham and Women's Hospital has been obtained for preliminary testing of the model, with filing and approval from the Institutional Review Board (approved protocol attached).

b. Interim analysis of the system using abstracted cases will be conducted

These cases are being constructed based on actual retrospective data collected from the Brigham and Women's Hospital. We anticipate having a first set of abstracted cases for formal evaluation (comparing performance of the computer-based system against that of oncologists) in September, 1999.

c. System will be refined in terms of belief network model and GUI given interim analysis results and internal user feedback.

Belief network models are still preliminary. We have concentrated on solving the technical problems, while oncologists are starting to build first networks representing eligibility criteria dependencies.

The following tasks are anticipated for Months 16-24, and the plan has not been modified from the one stated in the Approved Statement of Work:

Task 5. **Subject recruitment, abstraction of medical records, and creation of survey instruments for final analysis, Months 16-24:**

a. Lay people ("patients") will be recruited

b. Medical records will be abstracted and randomized

c. On-line forms for recording selection of clinical trials for "patients" and providers will be built

d. Surveys for assessing "patient" and provider satisfaction with the system will be built

e. Primary care providers and oncologists will be scheduled for final experiments

Task 6. **Evaluation experiments, Months 25-33:**

a. Oncologists will assess system's performance

b. "Patients" will use the system and fill on-line forms and surveys

c. Primary care providers will use the system and fill on-line forms and surveys

Task 7. **Final analysis and report writing, Months 34-36:**

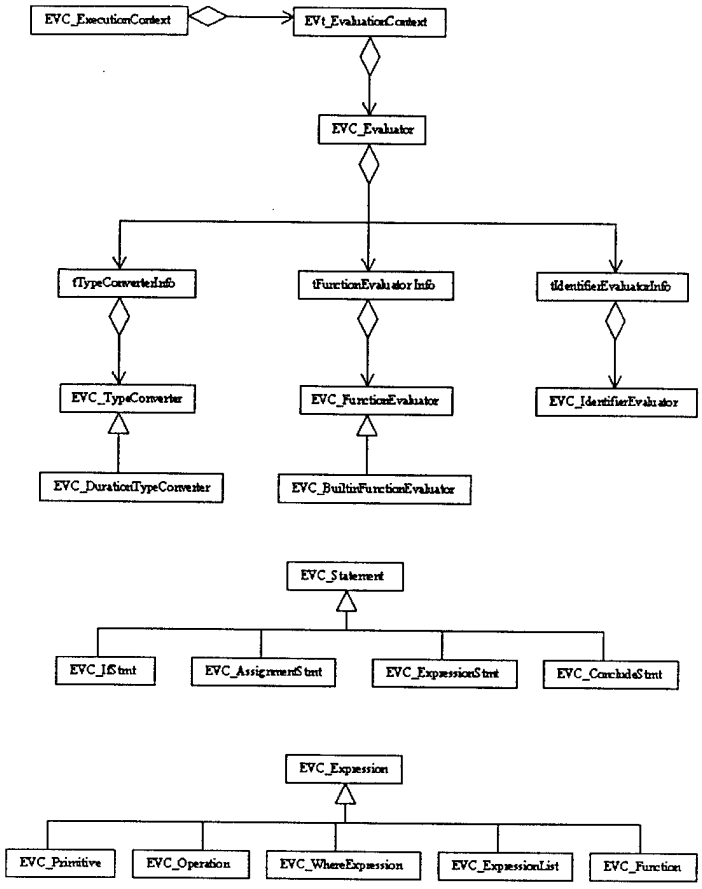
a. Final analyses of data from oncologists, "patients," and providers will be performed

c. A final report and manuscripts will be prepared

In the next sections, we describe the design of FACTs, and illustrate with some screen samples from the existing prototype.

Design

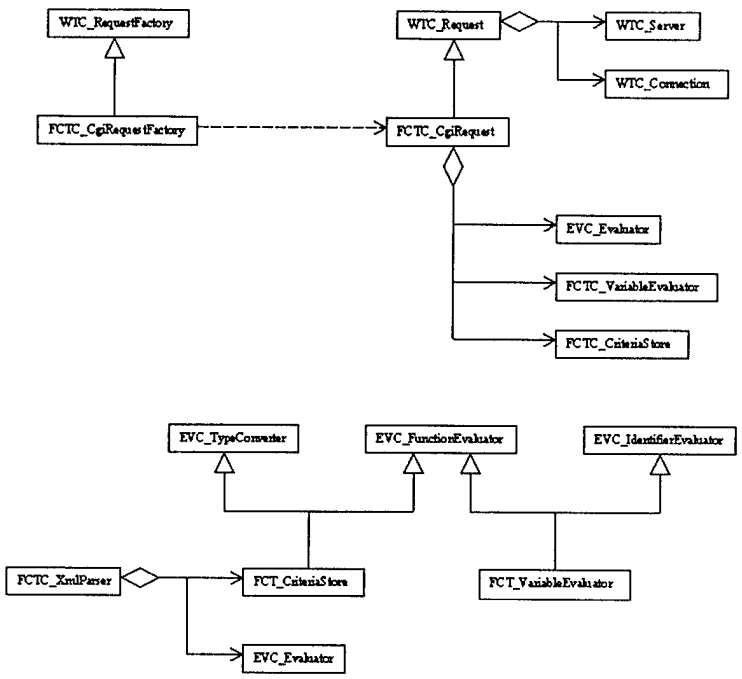
FACTS utilizes an evaluation engine called EV to interpret Arden statements and expressions, including logical and temporal criteria. EV uses a lexer generated with flex 2.5.4 and a parser generated with Bison 1.25. Information about the clinical trial protocols, including the encoded criteria, is stored in XML documents. A separate XML parser is used to obtain the portion of an XML document containing the criteria encoded in Arden. Then the EV parser constructs an abstract syntax tree representing Arden statements and expressions that can be interpreted by invoking its "Evaluate" method. The evaluation of the abstract syntax tree follows an interpreter design pattern to recursively request the objects representing the nodes of the tree to interpret themselves and yield the result of the evaluation. The following UML class diagram illustrates the object-oriented structure of EV. Statements and expressions are related by inheritance to allow their participation in an interpreter or visitor design pattern.



EV Project Class Diagram

12/4/98

The following UML class diagram illustrates the object-oriented structure of FACTS. The information in the encoded criteria is maintained by the criteria store object. Arden variables may be evaluated upon demand with the variable evaluator object. Identifier evaluators, function evaluators, and type converters may be registered with the EV evaluator object, which it will consider using in the course of evaluating a statement or expression.



FACTS Project Class Diagram

12/4/98

With regard to evaluating functions, EV provides a base class called EVC_FunctionEvaluator that may be derived from in other projects. These may be registered with EV to be potentially used in evaluation. In the EVC_Evaluator::EvaluateFunction method, the "Evaluate" methods of evaluators pointed to by elements in fFunctionEvaluatorSeq are invoked, starting with the last EVC_FunctionEvaluator that was registered and working backward, until an evaluator is found that does not yield an unknown error or the first evaluator in the sequence is reached. If a suitable evaluator is found, its return value is returned. If no suitable evaluator is found, this method yields a lookup error.

With regard to obtaining values of identifiers, EV provides a base class called EVC_IdentifierEvaluator that may be derived from in other projects. These may be registered with EV to be potentially used in evaluation. In the EVC_Evaluator::GetIdentifierValue method, if an identifier is not known in the immediate context, the identifier evaluators in fIdentifierEvaluatorSeq are searched in reverse order, beginning with the last one to be registered. A particular evaluator is asked to determine the value of the identifier by calling the Evaluate method. If no error is generated, the identifier is considered to have been found. If there was an unknown error or lookup error, then searching continues. If there was another type of error, the routine fails. If after these lookup attempts the identifier has still not been found, the routine signals an lookup error.

With regard to evaluating sentences in DSG Arden in the form of an abstract syntax tree, EV mostly uses the interpreter design pattern. Work has been done on extended the capabilities of EV to use alternative evaluators for "where" expressions. The visitor design pattern is being implemented to accomplish this.

Development Retrospective

The FACTS project was initially developed to run on a UNIX server. Subsequently, it was modified to run on a Windows NT server. In 11/98, the variable counting algorithm used in FACTS was modified slightly. The variable counting algorithm in FCTC_CgiRequest::TallyVars was formerly the following.

The score for a particular variable is the number of criteria that are not definitely known that the variable appears in over the protocols that have not been probably ruled out or definitely ruled out.

This algorithm was changed to the following.

The score for a particular variable is the number of protocols that: 1) have not been probably ruled out or definitely ruled out and 2) the variable appears in within a criterion that is not definitely known.

The former algorithm allowed a particular variable to be counted multiple times in one protocol, whereas the new algorithm limits the count for a particular variable to one per protocol. The two different algorithms can produce different results, especially when a protocol specifies a variable in multiple criteria. The most salient example of this that was found during testing involves the variable "metastases_locations". There are some protocols in which this variable occurs several times.

In 11/98, a variable constraining strength algorithm was implemented. As a preliminary measure, the variable ranking algorithm was refined to include a constant weight for each variable. The weight should satisfy

$0 \leq \text{weight} \leq 1$

and may be included in the XML file where any particular variable is described. The default variable weight is unity. The basic notion behind the weight is that it is the overlap between subpopulation prevalence and protocol disqualification.

The measure of the degree to which an unknown variable has the potential to rule out additional protocols may be called the "rule-out power" of the variable or the "constraining strength" of the variable (how strongly the variable constrains the set of operative protocols). The constraining strength s_i of the ith variable is

$s_i = F_i * w_i$

where F_i is the frequency of the ith variable (the fraction of the protocols that have not been probably ruled out or totally ruled out that the variable appears in), and w_i is the weight of the ith variable.

Changes to FCTC_XmlParser::ParseVariables were made to incorporate the ability to parse variable weights. Changes to FCTC_CgiRequest::TallyVars were made to perform the computation of the constraining strength for each tallied variable. Changes to FCTC_CgiRequest::PrintResults were made to display the variables of interest in ranked order. Class definitions were augmented with additional data members and header files with additional type definitions as needed to track the additional information.

In 11/98, the server code was converted to an ActiveX object, and an Active Server Page was used to invoke the ActiveX object and dynamically generate the HTML document presented to the client as the result of a FACTS search.

In 12/98, some of the error handling was optimized for use under ActiveX. The reporting of warnings to the browser under the ActiveX object project has been enabled for the parts of the code that use FCTC_CgiRequest::fWarnings or FCTC_XmlParser::fWarnings, either directly or indirectly. Not all such handling of errors actually output warnings; some of the mechanisms used were incompatible with the recent change to ActiveX. The capability of the function ErrorText in the file FCT_Request.cpp has been expanded to explicitly handle several additional error codes. This should improve the specificity of reporting warnings.

In 1/99, some minor operator name changes were effected to increase compatibility with Arden. Formerly, the "and" operator had an alternative name "&&", and the "or" operator had an alternative name "||". This is no longer the case. Now the "and" operator has an alternative name "&", and the "or" operator has an alternative name "|". This was done to avoid conflicts with the Arden concatenation operator "||".

In 3/99, changes to the Arden interpreter were made to enable enhancements to the "where" operator in DSG Arden. Inheritance relationships among enum values was already supported previously in the FACTS code, and the "is-a" operator works on them. An allowance for parents of a FACTS data type (as opposed to value) was made at this time to enable inheritance relationships among struct fields.

The behavior of the "where" operator in FACTS has been modified so that the left argument of the "where" operator is expanded to include all hyponyms (descendants) of all items in that argument. If an item in the left argument is a member of a struct which has inheritance relationships to other structs, then the list formed from the left argument is expanded to include also the corresponding members of all hyponyms (descendants) of the struct.

There was a requirement that the code to accomplish this alternative interpretation of the "where" operator reside in the FCTL project and not the EV project. This has been done so that EV does not know about this code specifically but will call this code when appropriate. This involved changes primarily to FCTL and also to EV, but the changes to EV were basically for defining the base visitor class only. The default behavior of EV is unchanged. In other words, these changes to EV are backward compatible with previous versions. The actual use of EV is the same.

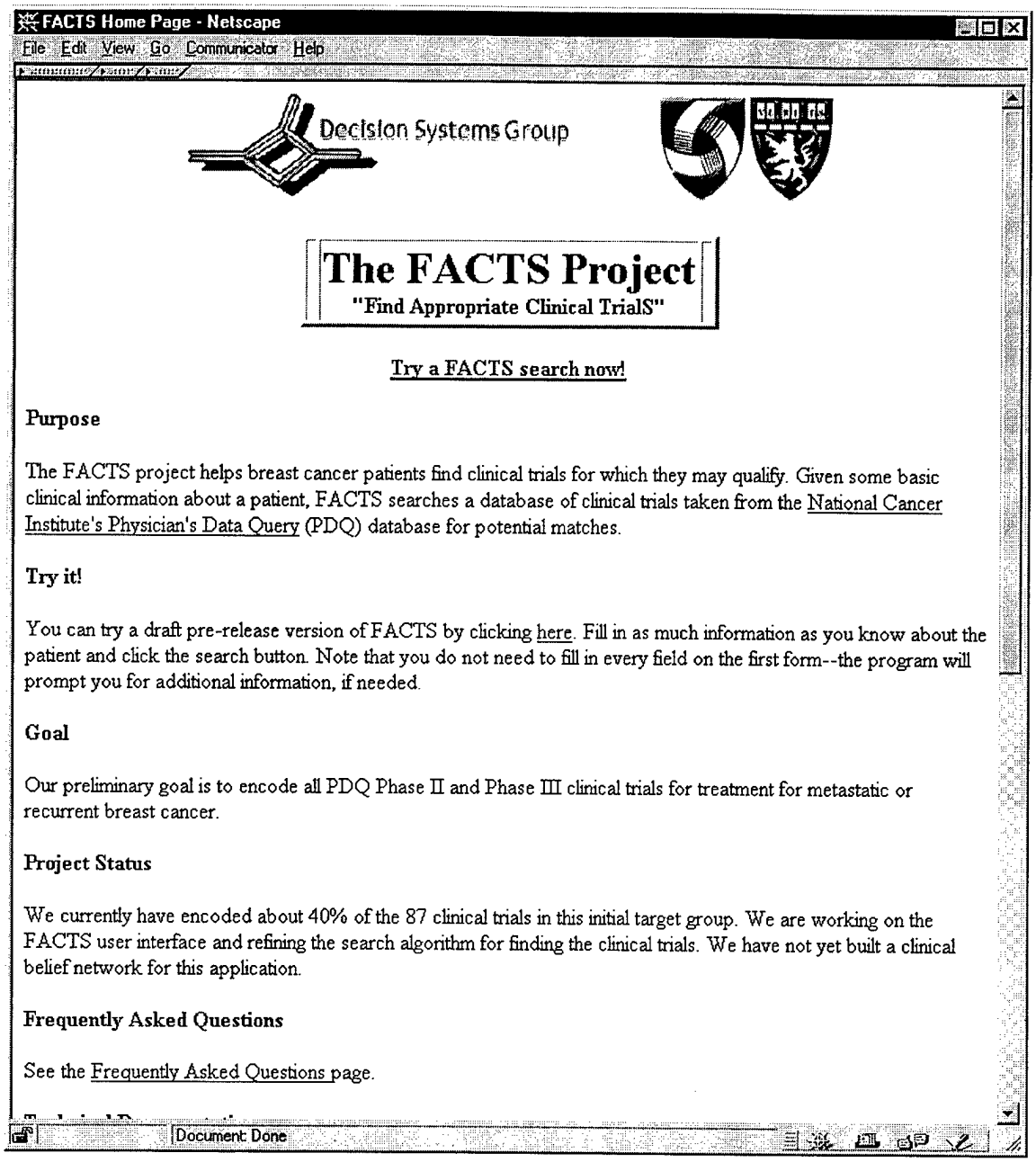
The mechanism by which the enhancements operate is essentially that the visitor design pattern is used instead of the interpreter design pattern, which was used in the relevant parts of EV previously. A pure visitor design pattern was not used because it would result in changing the interface to abstract syntax trees in EV, and hence existing code that uses EV could not be easily reconfigured to take advantage of new features of this type. The interface could also be expanded later if desired.

EV contains the base class for visitors. Derived visitors may be defined in other projects to provide alternative interpretations of DSG Arden abstract syntax trees constructed in EV representing DSG Arden sentences (statements or expressions). Access to the operative visitor (of base type EVC_Visitor) is controlled by a configurable singleton (of type EVC_VisitorSource). Basically, to use a different visitor that provides an alternative interpretation, you would only need to reconfigure the "visitor source" with your visitor. Then the rest of the code that uses EV can be used in the same way, but your visitor will be used for the interpretation instead. A visitor is configured by invoking the SetVisitor method on the EVC_VisitorSource object.

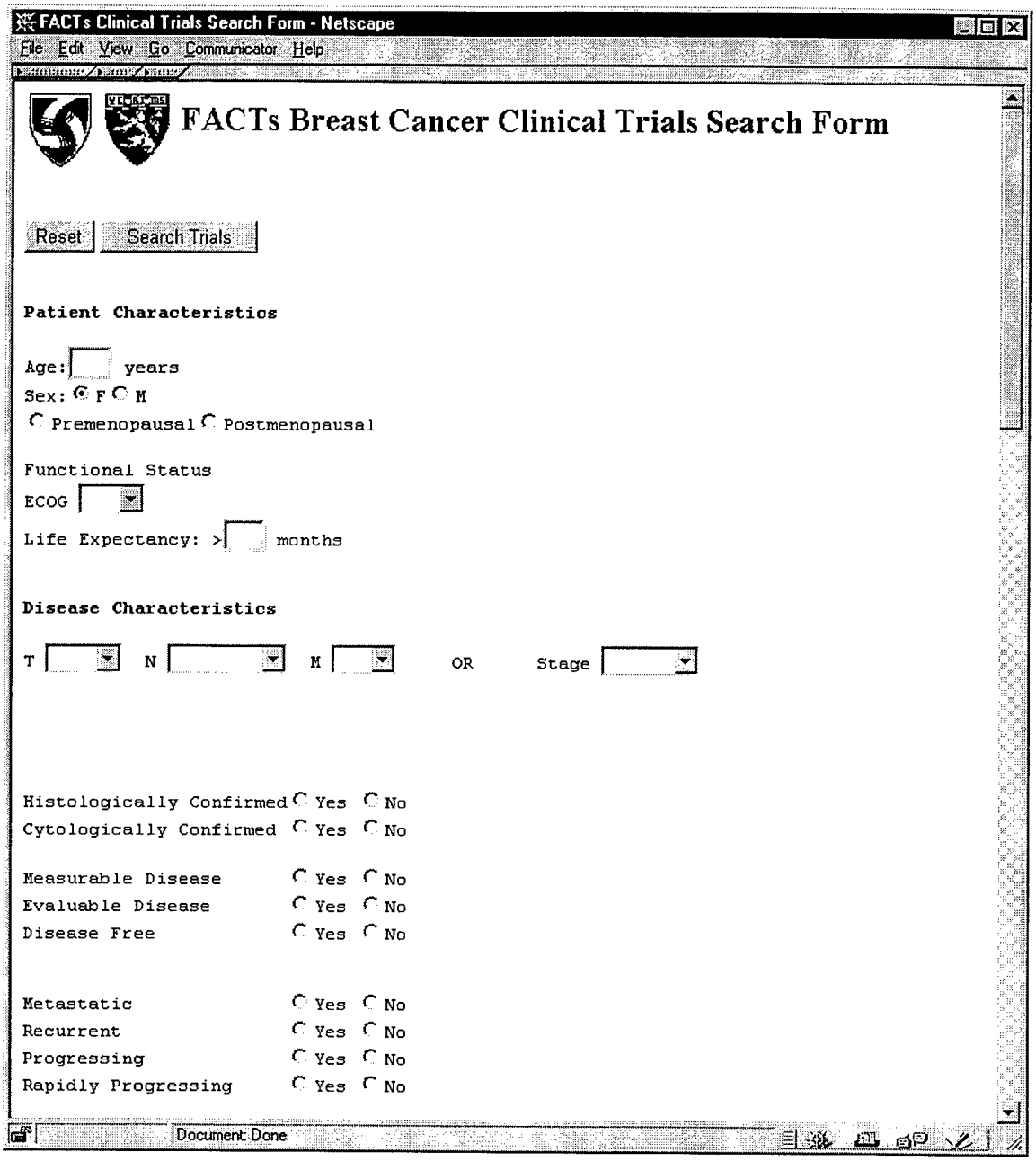
The EVC_Visitor class provides the basis for a visitor design pattern to interpret the DSG Arden abstract syntax tree constructed by EV. In the visitor design pattern, a node (generally an object) in an abstract syntax tree is interpreted (evaluated/executed) by an outside object (the visitor). This is in contrast to the interpreter design pattern, in which the node itself contains the interpretation logic. With the interpreter design pattern it is difficult to extend the way interpretation is done, because the application logic that accomplishes the interpretation is hard-coded into the nodes themselves. With the visitor design pattern, it is relatively easy to extend the way the nodes are interpreted; since the application logic that accomplishes the interpretation is put in a separate class, a new class can be derived that carries out the alternative interpretation.

The reason for using a visitor at this time is to allow a different interpretation for a "where" expression without putting the application logic for the alternative interpretation in the EV project. Specifically, the FACTS project has a need for interpreting "where" expressions in a way that makes use of information about inheritance relationships in the data model used in FACTS. In the future other projects can also implement their own interpretations by deriving a visitor subclass.

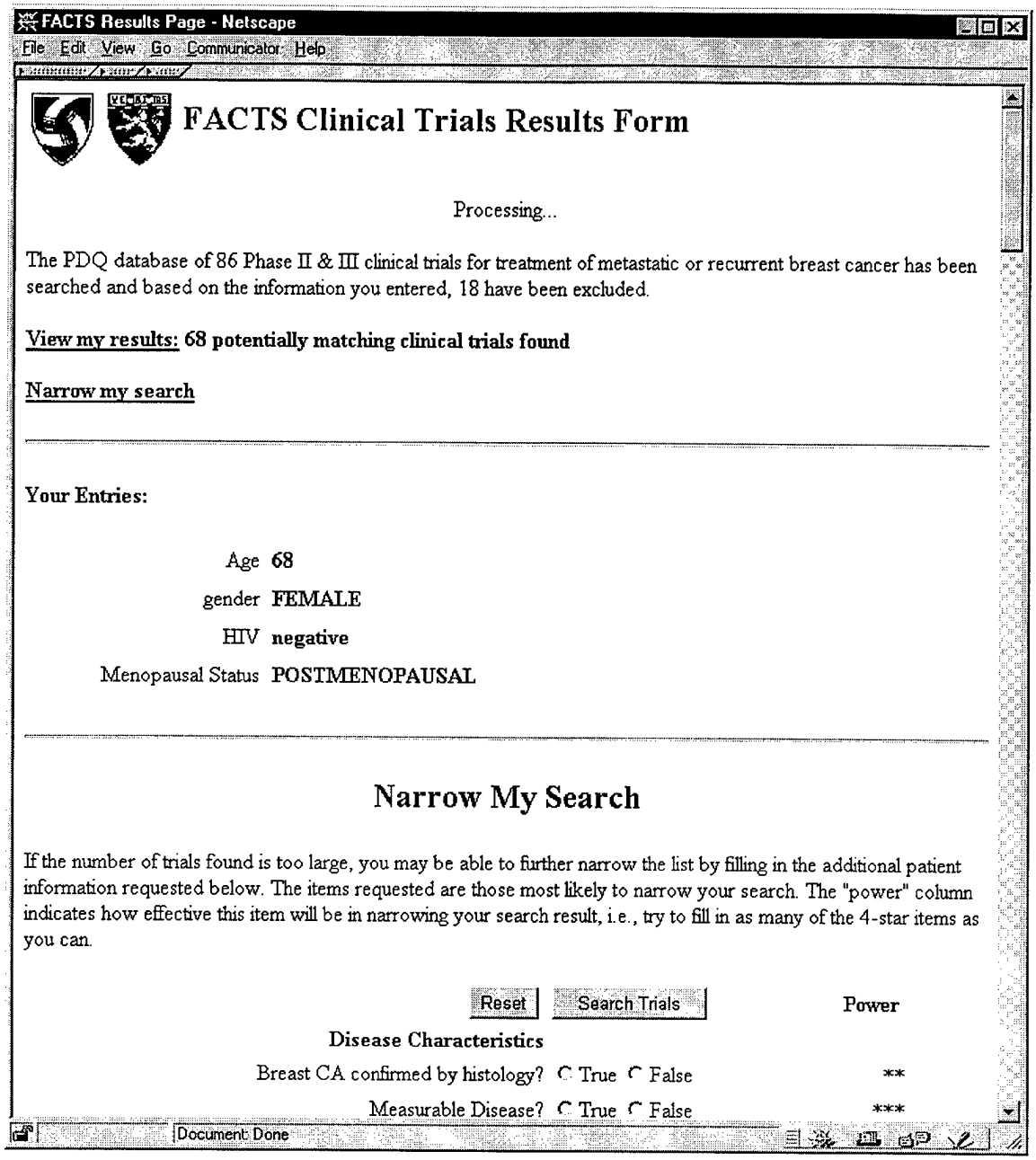
The graphical user interface for FACTS is also being enhanced. The following screenshot shows the old version of the FACTS home page.



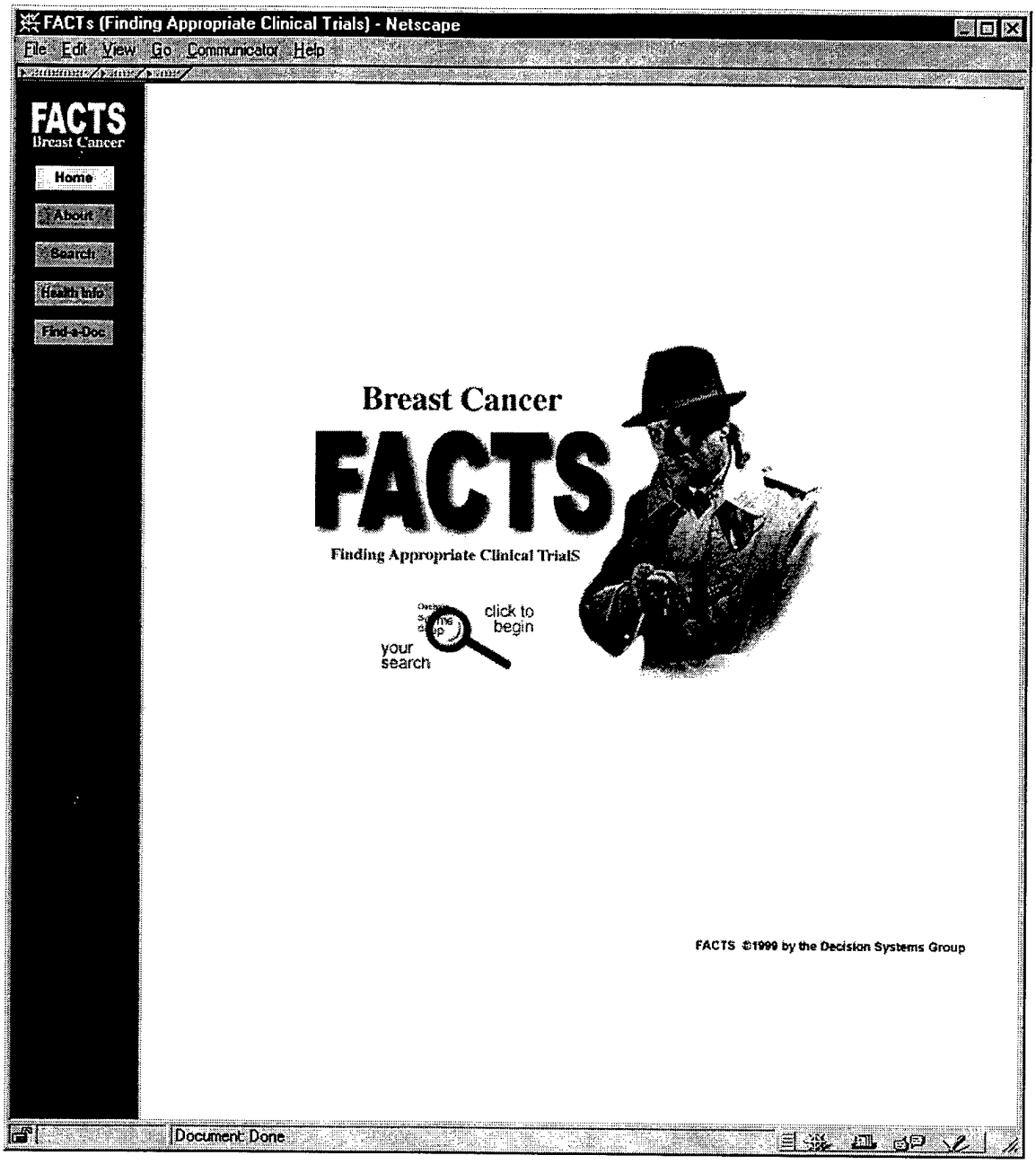
The following screenshot shows the old version of the FACTS search page.



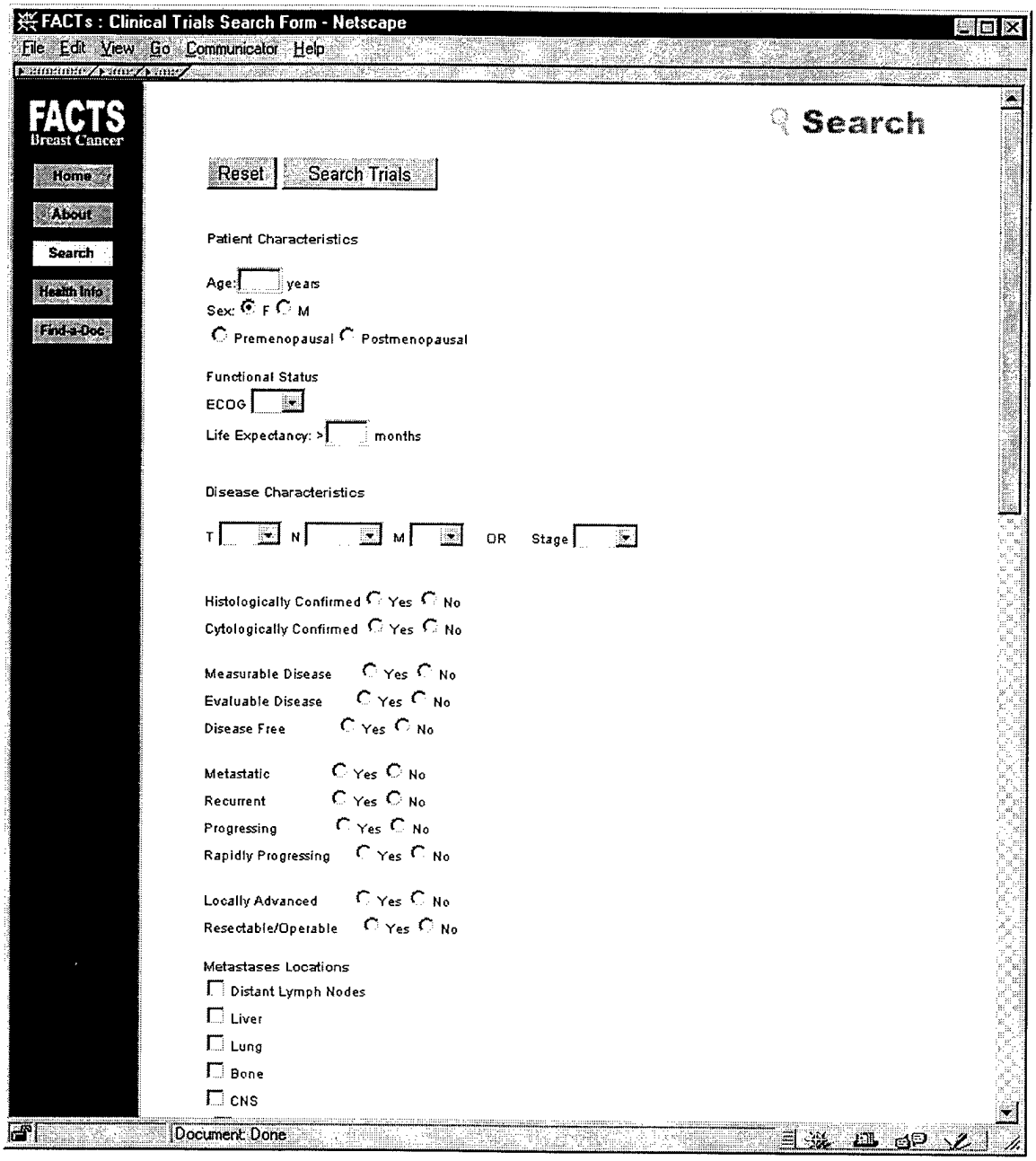
The following screenshot shows the old version of the FACTS search results page.



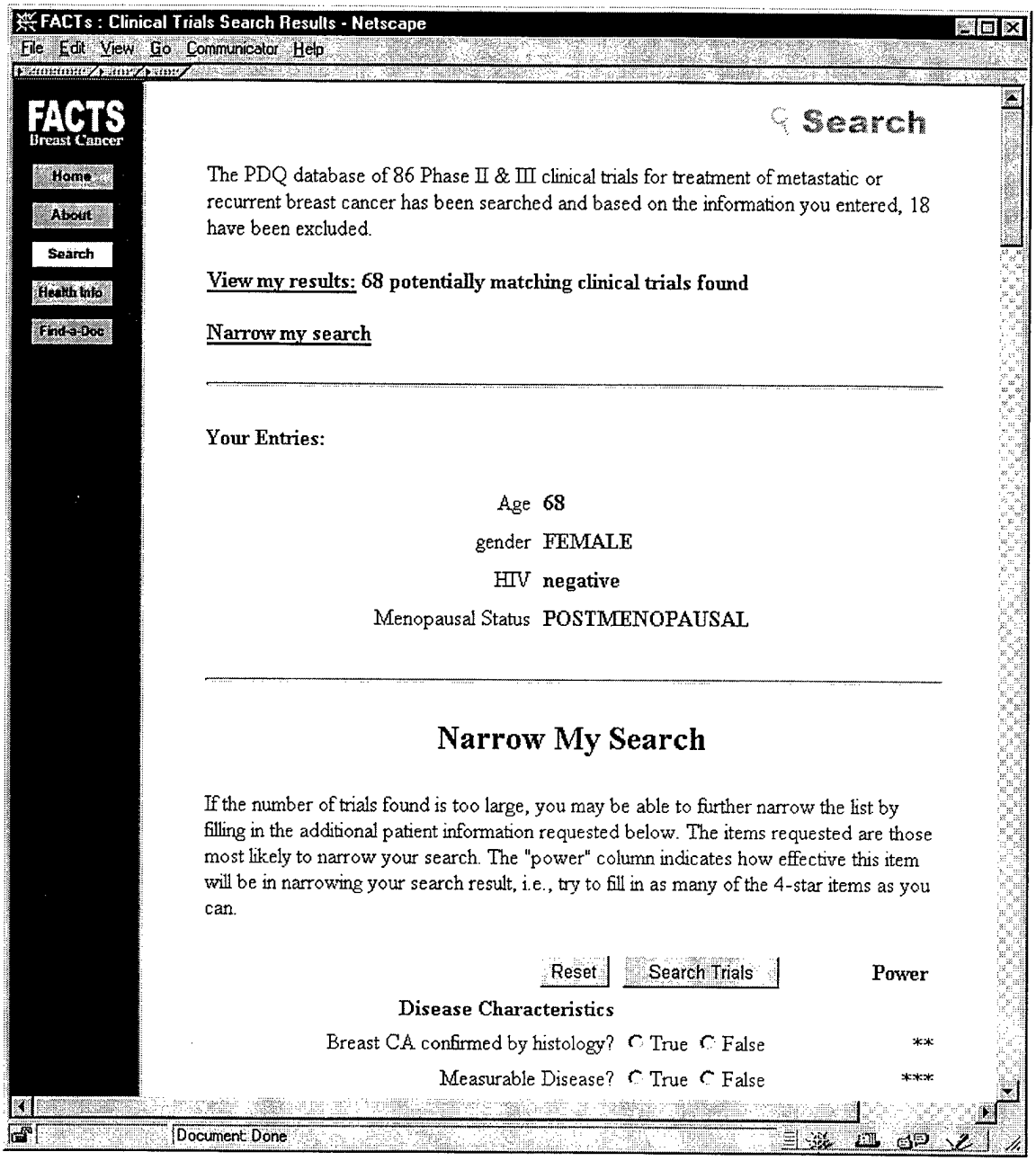
The remaining screenshots show the new versions of the FACTS Web pages under development. The following screenshot shows the new version of the FACTS home page.



The following screenshot shows the new version of the FACTS search page.



The following screenshot shows the new version of the FACTS search results page.


FACTs : Clinical Trials Search Results - Netscape
File Edit View Go Communicator Help
FACTS
Breast Cancer
Home
About
Search
Health Info
Find a Doc
Search
The PDQ database of 86 Phase II & III clinical trials for treatment of metastatic or recurrent breast cancer has been searched and based on the information you entered, 18 have been excluded.
View my results: 68 potentially matching clinical trials found
Narrow my search
Your Entries:
Age 68
gender FEMALE
HIV negative
Menopausal Status POSTMENOPAUSAL
Narrow My Search
If the number of trials found is too large, you may be able to further narrow the list by filling in the additional patient information requested below. The items requested are those most likely to narrow your search. The "power" column indicates how effective this item will be in narrowing your search result, i.e., try to fill in as many of the 4-star items as you can.
Reset
Search Trials
Power
Disease Characteristics
Breast CA confirmed by histology? True False **
Measurable Disease? True False ***
Document: Done

Key Research Accomplishments

- Isolated variables present in eligibility criteria for 85 protocols in PDQ
- Created and implemented structure for storing eligibility criteria and protocols
- Created syntax for representing eligibility criteria, based on modification of Arden syntax
- Implemented parser for extended Arden syntax
- Encoded 85 protocols using structure in XML and Arden syntax
- Implemented simplified patient data model
- Implemented graphical user interface to acquire patient data
- Developed deterministic engine to match patient values against eligibility criteria
- Developed ad-hoc algorithm to rank protocols in reverse order of appropriateness for a particular case
- Implemented graphical user interface to display summarized patient data
- Implemented graphical user interface to display appropriate protocols
- Implemented algorithm to select most informative variables for a given case
- Implemented graphical user interface to display most informative variables
- Started formative evaluation of system's performance
- Started graphical user interface refinement based on oncologist's recommendations
- Redesigned evaluation process
- Obtained approval from IRB to test system with abstracted cases from Brigham and Women's Hospital

Reportable Outcomes

Manuscripts

Ohno-Machado L, Boxwala AA, Wang SJ, Mar P. Decision Support for Clinical Trial Eligibility Determination in Breast Cancer. Technical Report TR-199-02, Decision Systems Group. Submitted to the 23rd Annual Americal Medical Informatics Association Fall Symposium.

Abstracts

Ohno-Machado L, Wang SW. Selection of Clinical Trials Using Artificial Intelligence. Abstract presented at the 1999 Breast Cancer Research Symposium of the Massachusetts Department of Public Health.

Wang SJ, Ohno-Machado L, Boxwala A, Greenes RA. Enhancing Arden Syntax for Clinical Trial Eligibility Criteria. .Technical Report TR-199-02, Decision Systems Group. Abstract Submitted to the 23rd Annual Americal Medical Informatics Association Fall Symposium.

Presentations

Poster presentation at the 1999 Breast Cancer Research Symposium of the Massachusetts Department of Public Health, 4/28/99.

Presentation at the Seminar for the Medical Decision Making Group at the Laboratory for Computer Science, Artificial Intelligence Labs, Department of Electrical Engineering and Computer Science, MIT, 12/8/98.

Informatics such as databases

Database of Encoded Protocols available at http://telmato.bwh.harvard.edu:8000/FACTS/data/

Conclusions

We have accomplished the first steps towards the construction of an automated system to determine patient eligibility for a breast cancer clinical trial. We have started by defining the necessary syntax and storing structures for data contained in clinical trial protocols and entered as patient values. We have implemented a parser for this syntax, a deterministic algorithm to match patients against trials, and an ad-hoc algorithm to rank order the
We have completed the tasks in the approved statement of work that were planned for Year 1 of this project. A working prototype of the system we envision as our final product is available at http://telmato.bwh.harvard.edu:8000/FACTS/FormIn.htm.

Our next steps are to modify the eligibility determination engine by adding functionality for cases that contain missing data, and to account for associations among criteria and among patient data values. By doing this, we will avoid "double counting" of criteria, and will be able to develop a ranking algorithm based on probabilistic principles.

References

Ohno-Machado L, Boxwala AA, Wang SJ, Mar P. Decision Support for Clinical Trial Eligibility Determination in Breast Cancer. Technical Report TR-199-02, Decision Systems Group. Submitted to the 23rd Annual Americal Medical Informatics Association Fall Symposium. [Appendix 1]

Wang SJ, Ohno-Machado L, Boxwala A, Greenes RA. Enhancing Arden Syntax for Clinical Trial Eligibility Criteria. .Technical Report TR-199-02, Decision Systems Group. Abstract Submitted to the 23rd Annual Americal Medical Informatics Association Fall Symposium. [Appendix 2]

Ohno-Machado L, Wang SW. Selection of Clinical Trials Using Artificial Intelligence. Abstract presented at the 1999 Breast Cancer Research Symposium of the Massachusetts Department of Public Health. [Appendix 3]

Appendix 1- Paper

Decision Support for Clinical Trial Eligibility Determination in Breast Cancer

Lucila Ohno-Machado, MD, PhD, Samuel J. Wang, MD, PhD,
Perry Mar, Aziz A. Boxwala, MBBS, PhD
Decision Systems Group and Health Science and Technology Division
Harvard Medical School and Massachusetts Institute of Technology, Boston, MA

Abstract

We have developed a system for clinical trial eligibility determination where patients or primary care providers can enter clinical information about a patient and obtain a ranked list of clinical trials for which the patient is likely to be eligible. We used clinical trial eligibility information from the National Cancer Institute's Physician Data Query (PDQ) database. We translated each free-text eligibility criterion into a machine executable statement using a derivation of the Arden Syntax. Clinical trial protocols were then structured as collections of these eligibility criteria using XML. The application compares the entered patient information against each of the eligibility criteria and returns a numerical score. Results are displayed in order of likelihood of match. We have tested our system using all phase II and III clinical trials for treatment of metastatic breast cancer found in the PDQ database. Preliminary results are encouraging.

Introduction

Historically, accrual of patients for clinical trials has not been very successful, particularly for certain clinical domains. Studies demonstrate that just a small percentage of eligible patients (3 to 10%) are actually enrolled in such trials [1,2]. The low accrual rates are attributed to: (1) physician factors such as lack of knowledge about clinical trials, (2) patient factors such as lack of patient-oriented information regarding trials, (3) organizational barriers, and (4) health care system obstacles. If clinical trial information can be made more accessible to patients and their primary care providers (PCPs), we believe that clinical trial accrual rates can improve.

The increasing participation of patients in decisions regarding their own health has created a demand for health information resources oriented towards the patient and PCP, rather than the specialist [5]. A few systems have been previously designed to help with the determination of clinical trial eligibility. Tu et al. developed systems for this purpose, described in [6]. Ohno-Machado et al. previously developed a system that could reason under conditions of uncertainty [7]. However, these systems have focused on helping investigators identify eligible patients for a specific clinical trial. In contrast to these systems, the purpose of our system is to enable PCPs and patients to identify the best trials for a specific patient.

Material and Methods

Data. We used the National Cancer Institute's Physician Data Query (PDQ) database [8] as the source of information for clinical trials. The clinical trial summaries in the PDQ database contain free-text lists of eligibility criteria organized by patient characteristics (e.g., age, menopausal status); disease characteristics (e.g., histology, metastases); and prior and concurrent therapy. For the preliminary phase of this study, we selected from the PDQ database all Phase II and Phase III trials for the treatment of metastatic or recurrent breast cancer. Breast cancer was chosen because this is the oncology domain that contains the largest number of clinical trials. We chose advanced stage cancer because we hypothesized that these patients would be more interested in seeking participation in clinical trials after exhausting traditional treatment venues. We decided to limit our initial set to Phase II and Phase III trials since these studies are further developed, and typically involve several patients. We found a total of 85 clinical trials in the PDQ database (as of July 1998) that fit these parameters.

Clinical Trial Eligibility Database. Each clinical trial summary was encoded into a structured format. The encoded summary was stored in an XML document (Figure 1). This document contains elements describing identifying information about the clinical trial (name of trial, protocol number) and a collection of criteria elements. Each criterion element contains the original narrative text description from PDQ and the criterion encoded in a computable expression. The criterion is encoded in a modified version of the grammar used for specifying logic statements in the Arden Syntax [9]. Modifications had to be made to the Arden Syntax specification in order to accommodate a data model that contains hierarchical term relationships and compound data-types. (Details and discussion of our modifications to the Arden Syntax are presented elsewhere [10].) The resulting extended syntax for conditional expressions is also being incorporated into proposed extensions to GLIF, a clinical guideline interchange format developed by The InterMed Collaboratory [11].

```
<PROTOCOL ID="09251">
 <NAME> Phase II Randomized Study of
Cyclophosphamide/Methotrexate/Fluorouraci
l (CMF)
vs Mitoxantrone in Elderly Patients with
Advanced Breast Cancer
 </NAME>

<!--Disease Characteristics-->
<CRITERION>
  No CNS metastases
  <SPEC>
    (metastases_locations where it is a
"CNS") == []
  </SPEC>
</CRITERION>

<!--Patient Characteristics-->
<CRITERION>
 Over 70
  <SPEC>
    age &gt; 70
  </SPEC>
</CRITERION>

<CRITERION>
 Postmenopausal
   <SPEC>
     menopausal_status ==
"postmenopausal"
   </SPEC>
</CRITERION>

<!--Hematopoietic:-->
<CRITERION>
 WBC at least 3,000
   <SPEC>
    WBC &gt;= 3000
   </SPEC>
 </CRITERION>
```

Figure 1. Excerpt of clinical trial protocol structured in XML format.

The translation of the original free-text criterion descriptions from PDQ into a machine-interpretable representation was largely a manual process performed by informatics fellows and faculty in our laboratory. We used text parsing tools such as Perl scripts to automate portions of this process. We established a uniform basis for encoding criteria. For example, a certain clinical trial summary may have specified "estrogen receptor negative," and another may have specified "ER (-)." These refer to the same eligibility criterion and are encoded using the same expression ("estrogen_receptor == negative").

```
<!-- Patient Characteristics -->

<VARIABLE NAME='age' TYPE='number'
CUI='C0001779'>
</VARIABLE>
```

```
<VARIABLE NAME='birthdate' TYPE='date'
CUI='C0421451'>
</VARIABLE>

<VARIABLE NAME='gender' TYPE='enum'
CUI='C0079399'>
Gender of patient
  <VALUE CUI='C0024554'>male</VALUE>
  <VALUE CUI='C0015780'>female</VALUE>
</VARIABLE>

<VARIABLE NAME='menopausal_status'
TYPE='enum' CUI='C0025320'>
Menopausal status of patient.
  <VALUE
CUI='C0279752'>premenopausal</VALUE>
  <VALUE
CUI='C0279753'>postmenopausal</VALUE>
</VARIABLE>
```

Figure 2. Excerpts from data dictionary containing definitions of clinical concepts used in the eligibility criteria.

In order to adequately model eligibility criteria, we found it necessary to create a data model that was sophisticated enough to accommodate hierarchical relationships among clinical concepts, sub-attributes of concepts, and temporal relationships among concepts. The concepts used in the eligibility criteria were defined in a data dictionary (also an XML document) (Figure 2), and mapped to concepts in the UMLS Metathesaurus [12]. We analyzed all the encoded criteria to assess which concepts occurred most frequently and were also relatively easy for the patient or PCP to obtain. This information was taken into consideration to construct web-based entry forms, shown in Figure 3.

Clinical Trial Ranking. Upon entry of patient data, the application produces a ranked list of clinical trials that the patient is eligible for. The ranking algorithm is tolerant of missing data. All criteria are considered as having equal weight (importance) when used in protocol ranking. The algorithm sequentially processes all the criteria in all the clinical trials. The algorithm first rules out all clinical trials for which at least one eligibility criterion was not met. For the remaining clinical trials, the ones that have fewest unknown criteria are placed higher on the list. Resulting trials are displayed with links to the original PDQ clinical trial summaries (Figure 4). The search can be refined with data entered in dynamically created forms (Figure 5). For each clinical trial, we also provide a summary of which criteria have been met and which still need to be evaluated (Figure 6).

Application. We are developing two versions of the application: one for the primary care provider and one for the patient. The version for the patient will provide a simplified user interface and will only request data

that a patient would be expected to know. The application runs on the Microsoft Windows platform. HTML pages are dynamically generated on the server using Microsoft's Active Server Pages (ASP). The application logic was written in Visual C++ and wrapped as an ActiveX object that is invoked by ASP.

Results

A total of 2188 criteria in the set of 85 clinical trials were chosen for this study. In this set, the least, most, and median number of criteria in a protocol were 6, 45, and 25 respectively. To date, we have encoded about 50% of the criteria in these clinical trials. We are first encoding frequently occurring criteria and those that are readily accommodated by the criteria representation syntax. (See [10] for details on difficulties encountered in encoding the eligibility criteria.) Figures 3 to 6 show an example of the PCP version of the application for a sample breast cancer patient: a premenopausal, 55 year-old woman with stage IV breast cancer with metastases to liver and bone, previous mastectomy, chemotherapy and radiotherapy. This patient also suffers from coronary artery disease and diabetes mellitus. Figure 3 shows the initial data input form in which the PCP has entered some clinical information about the patient. Using this information, the program returns a preliminary list of trials. This list is ranked, with the most likely matches at the top (Figure 4).

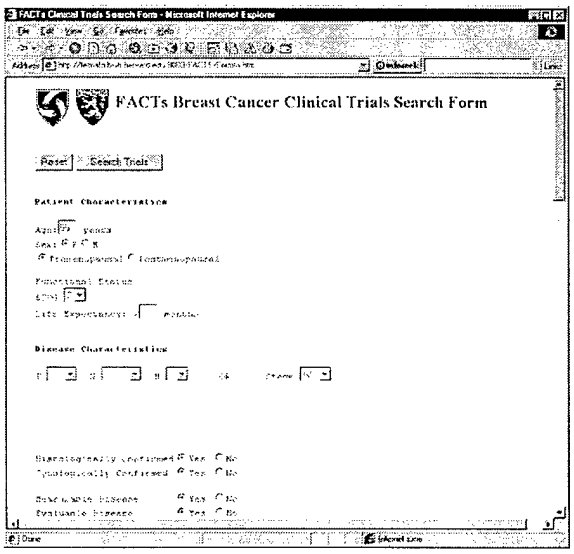


Figure 3. The initial entry form requests items that are most frequent and easiest to obtain.

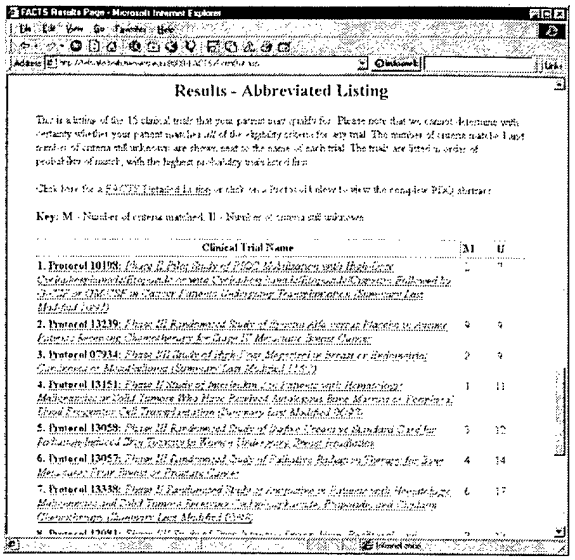


Figure 4. Results page showing a ranked list of clinical trials.

If the list is long, the application offers the PCP an opportunity to fill in additional patient information to narrow the search. The program dynamically constructs the secondary input form to request the information that would be more likely to narrow the number of clinical trials (Figure 5). Again, the PCP fills in as much additional information as he or she can. This process can be repeated as many times as desired until either the resulting list is short enough, or there is no additional information required or available.

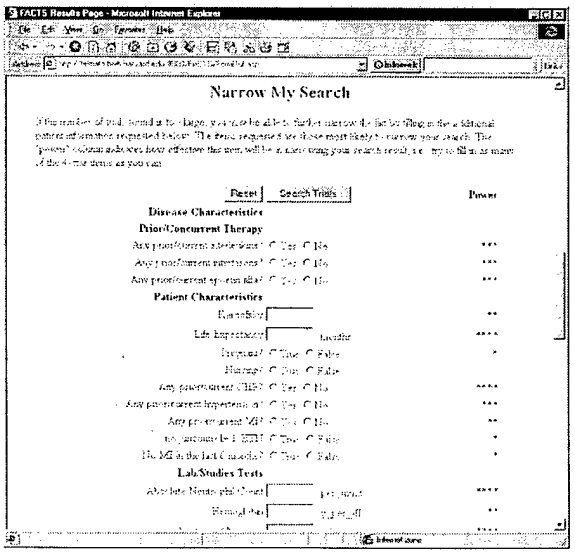


Figure 5. Secondary entry forms are created dynamically and request information that will be most useful in narrowing the search.

The final list is presented in order of likelihood of match. In this example, the system narrowed the list to 15 trials that the patient is potentially eligible for. A summary of all the entered information is provided.

Detailed information about these clinical trials (Figure 6) can be displayed, along with a list of the criteria still to be checked.

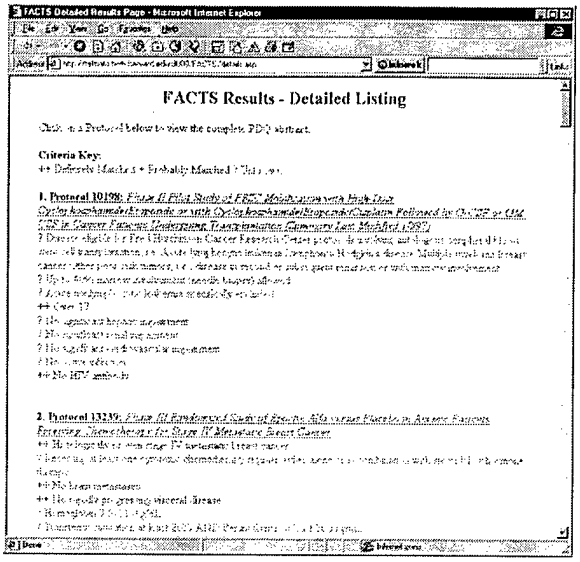


Figure 6. Detailed information about remaining trials is displayed.

Discussion and Future Directions

The current ranking algorithm makes two simplifying assumptions: (1) all criteria have equal importance and equal probability of being met if their values are unknown, and (2) all criteria are independent. Regarding the first assumption, a more accurate approach would be to assign a weight to each criterion or data item, and then use these weights to compute the ranking. We may be able to obtain these weights by asking domain experts, from the literature, or by analysis of large patient data sets. Tu [6] has proposed that some criteria variables are mutable over time (e.g., age) or controllable (e.g., stop current chemotherapy), and therefore might bear less weight in ruling-out or ranking one clinical trial against others. We have not decomposed criteria into "atomic" parts, each containing just one variable, hence this approach has not been yet tested.

The other simplifying assumption, criteria (and data item) independence, also introduces inaccuracies in ranking. For example, a clinical trial may specify two separate data items for the liver function tests, AST and ALT: "AST < 2 times normal" and "ALT < 2 times normal." These criteria are currently considered independent, when in fact a better approximation would be to consider them just *conditionally* independent given a certain liver disease. For example, if AST is high, there is an increased probability that ALT is high because the disease that causes the former to increase is also likely to cause the latter to do so. The independence assumption causes some criteria to be unfairly "counted twice." A more accurate approach would be to identify dependencies among the data items and adjust the scoring accordingly. In this version of the application, we considered all criteria to be Boolean (i.e., "true" or "false"), and have not further characterized their nature.

The current clinical trial selection algorithm is deterministic. We have not attempted to deal with uncertainty using probabilities in this prototype. A global model to infer the value of missing values for common criteria and specification of criteria dependencies will be built using expert knowledge. This model will be based on a belief network, the structure and probabilities of which will be extracted by interviews with specialists, analysis of literature, or "learned" from clinical databases. A future version of this system will take into account "proxies" for certain criteria (e.g., known renal disease as a proxy for laboratory values that measure renal function, or "severity of cancer" as a proxy for staging). The probabilities of eligibility will be determined by inferencing values for required data from the proxies.

Other prototype applications have been built with the assumption that certain medical domains may require very few eligibility criteria to reasonably eliminate a large percentage of the candidate trials for a given patient [13]. In contrast, our approach has been to attempt to encode as many criteria as we reasonably can in an attempt to arrive at a more accurate list of potentially matching clinical trials. However, it is difficult to algorithmically determine eligibility with 100% accuracy because of the clinical judgement that is necessary for evaluating several of these criteria. Our objective is to narrow and rank the list of matching trials, as much as possible, before turning the list over to a specialist for final determination of eligibility. Encoding complex criteria is a time-consuming effort. Although we have developed some automated parsing tools to facilitate this task, it remains a largely manual process. We predict that our application will perform better as we encode more criteria. However, an open question that deserves further study is how much encoding is "enough," i.e., at what point is it not cost-beneficial to encode more complex criteria. Since software applications cannot determine clinical trial eligibility with 100% certainty, it may not be worth the extra effort to encode very complex criteria.

The criteria encoded for this study were taken from clinical trial summaries from PDQ. These summaries are abstracted from the original protocol documents and may lose some fidelity in the process. Our encoding is only as good as the translated text descriptions. For improving accuracy, an alternative approach would be to go directly to the original full research clinical trial descriptions to obtain the eligibility criteria. The future development and routine

use of computer-based protocol authoring tools may reduce these problems.

Currently, we have not taken into account patient preferences in ranking the clinical trials, such as modality of treatment, potential toxicity, potential for cure, and geographic constraints. The system currently ranks trials solely based on the likelihood that the patient will satisfy the eligibility requirements. It is a very different question to ask what types of trials a patient may prefer. While eligibility criteria are obviously a firm prerequisite to enrollment, in cases with incomplete information, there may be some benefit to introducing patient preferences even before eligibility has been completely determined. This could help narrow the list more quickly so as not to waste the patient's or clinician's time in reviewing eligibility requirements for trials that the patient would never consider enrolling in.

We plan to automatically retrieve some of the required patient data from the clinical information system at our institution in order to ease the data entry burden on the user. The user will only need to provide information not available in the clinical system. For the institutional version, we will link the eligibility component to other tools that automate the enrollment process, such as display of informed consent forms, and detailed explanation of the clinical trials. A more general version of the application will be available on the WWW. In addition to UMLS, we also plan to map the concepts used in our system to the Common Data Elements (CDE) that are being developed under the supervision of the informatics group at the National Cancer Institute [14]. Mapping to the CDE will make the system more robust for national scale use. The open architecture and facility to add customized dictionaries will also make it easy to adapt the system for integration to electronic medical record systems of different institutions.

This initial version of the application has been designed for use by PCPs. For the patient version, we intend to customize the user interface according to different levels of user sophistication. The user interface will be designed in consultation with patient advocacy groups, health educators, and PCPs. Reduction and simplification of data items to be entered is necessary. We will utilize a decision analytic approach to determine the data items needed.

Conclusions

We have developed a WWW-based decision support system to help patients and providers determine the patient's eligibility for certain clinical trials. The system currently contains all Phase II and III treatment clinical trials for metastatic breast cancer from the NCI's PDQ database. It rules out trials that the patients are not eligible for and ranks the remaining trials according to how many criteria still need to be checked to determine eligibility. This initial prototype system has helped us identify relevant issues in machine-readable criteria representation, user interface design, and clinical trial ranking under uncertainty. Preliminary testing of the system with a few clinical cases has been promising. A formal evaluation of usability and reliability is underway. Future versions of this application will include a belief network that will allow the system to impute missing data values and reason under conditions of uncertainty.

Acknowledgments

This project was funded by contract 34078PP1024 from the Massachusetts Department of Public Health and grant DAMD17-98-1-8093 from the Department of the Army. The contents of this article do not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Dr. Wang was funded by NLM grant2 T15 LM07092. We would like to thank Jeremy Theal, Dr. Jeff Huhn and Dr. Ross Martin for helping to encode the eligibility criteria. We also thank Dr. UrsulaMatulonis, Dr. Craig Bunnel, and Dr. Darrel Smith for sharing their expertise in breast cancer. Mr. John Ehresman implemented parts of this work. Prof. Robert Greenes provided valuable suggestions to this project.

References

[1] Mansour EG. Barriers to clinical trials. Part III: Knowledge and attitudes of health care providers. *Cancer*, 1994, 74(9 Suppl):2672-5.

[2] Winn RJ. Obstacles to the accrual of patients to clinical trials in the community setting. *Seminars in Oncology*, 1994, 21(4 Suppl 7):112-7.

[3] Taylor KM; Margolese RG; Soskolne CL. Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer. *New England Journal of Medicine*, 1984, 310(21):1363-7.

[4] Klabunde C; Kaluzny A; Ford L. Community Clinical Oncology Program participation in the Breast Cancer Prevention Trial: factors affecting accrual. *Cancer Epidemiology, Biomarkers and Prevention*, 1995, 4(7):783-9.

[5] Sweeney MA; Skiba D. Combining telecommunications and interactive multimedia health information on the electronic superhighway. *Medinfo*, 1995, 8 Pt 2:1524-7.

[6] Tu SW; Kemper CA; Lane NM; Carlson RW; Musen MA. A methodology for determining patients' eligibility for clinical trials. *Methods of Information in Medicine*, 1993, 32(4):317-25.

[7] Ohno-Machado L, Parra E, Henry SB, Tu SW, Musen MA. AIDS2: A decision-support tool for decreasing physicians' uncertainty regarding patients eligibility for HIV treatment protocols. *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care*, 429-33, 1993.

[8] NCI. http://cancernet.nci.nih.gov/.

[9] American Society for Testing and Materials. E 1460 Standard Specification for Defining And Sharing Modular Health Knowledge Bases (Arden Syntax for Medical Logic Modules). ASTM Standards, v 14.01. Philadelphia: ASTM, 1992; 539-87.

[10] Wang SJ, Ohno-Machado L, Mar P. Representing Criteria in Guidelines and Clinical Trial Protocols: Common Needs and Solutions. *DSG Technical Report* 1999-04.

[11] Ohno-Machado L, Gennari JH, Murphy SN, et al. The guideline interchange format: a model for representing guidelines. *J Am Med Inform Assoc*. 1998;5:357-72.

[12] Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc 1993;81(2):170-7.

[13] Gennari J. personal communication, 1998.

[14] NCI Cancer Informatics Infrastructure Home Page. http://hiip-wkst

Appendix 2 - Poster

Enhancing Arden Syntax for Clinical Trial Eligibility Criteria

Samuel J. Wang, MD, PhD, Lucila Ohno-Machado, MD, PhD, Perry Mar,
Aziz A. Boxwala, MBBS, PhD, Robert A. Greenes, MD, PhD
Decision Systems Group, Harvard Medical School, Boston, MA

Background. We have designed a general expression syntax based on the Arden Syntax[1] for representation of eligibility criteria in clinical trials. The goal of this project was to design a syntax to represent eligibility criteria for breast cancer clinical trial protocols. We are developing an application[2] that finds clinical trials for which a patient is eligible. The system compares the information entered by the user against each of the eligibility criteria in the clinical trials in the database and returns a rank-ordered list of matches. Most existing Arden applications do not address the more complex needs required to represent clinical trial eligibility criteria, such as representing medical problems and treatment data. Because of these requirements, we found it necessary to make extensions to Arden for our application. This enhanced Arden-based syntax could be adopted as the standard for representing logic for the GuideLine Interchange Format (GLIF), an object-based guideline representation language proposed by the InterMED collaboratory[3].

Approach. Below is a list of the enhancements we made to the Arden Syntax.

We introduced the concept of an enumerated data type where variables can only be assigned previously declared enumerated string values. For example, we constrained the variable "menopausal_status" to only be able to take on the values "premenopausal" or "postmenopausal." By using enumerated data types, we can map synonyms or spelling variants to controlled terms at compile time.

Representation of a disease condition or prior treatment often requires specific attributes associated with that particular concept denoting severity, status, type, or location. To accommodate this need, we introduced the concept of a "struct" data type and a corresponding "dot operator" (e.g., "CHF.severity") that allows us to assign specific attributes such as severity or status to each variable.

Many clinical concepts are hierarchical by nature, such as diseases, medications, and treatments. In addition, clinical data are often represented in varying degrees of specificity. For example, the language needs to be able to determine that "congestive heart failure" *is a* "cardiovascular disease." Also, because of the complex nature of disease classification, it is necessary for us to be able to accommodate multiple inheritance of a given disease. Thus, we have designed our system to allow construction of these types of directed acyclic inheritance graphs.

Arden has the concept of a single primary time associated with data variables. While this works well for "instantaneous" clinical data types such as laboratory test results, other clinical concepts occur over intervals of time, such as diseases and treatments. We enhanced Arden's time handling to allow representation of time intervals by introducing both a "start time" and an "end time" for all variables. This structure enables us to explicitly define the time interval of a patient's diseases and treatments. This allows us to form expressions such as "start_date(recurrence) > end_date(chemotherapy) + 4 weeks".

The choice of clinical vocabulary is tightly linked to the implementation of a practical data-query and data-modeling scheme for any Arden implementation. Clinical vocabulary terms must be controlled at design time in order to obtain consistent query results. For our application, we used the UMLS Metathesaurus to allow the greatest flexibility in the choice of terms available, while still maintaining control over the mapping of synonyms.

Application. For the clinical trial eligibility project, we built a database from information taken from the NCI's Physician Data Query (PDQ) database. We retrieved all phase II and phase III clinical trial summaries pertaining to treatment of metastatic breast cancer (as of July, 1998). We are encoding the resulting 2188 criteria from these 85 protocols using our enhanced version of the Arden syntax. We estimate that approximately 90% of the eligibility criteria can be encodedusing our new expression syntax and data model.

Conclusion. We have developed and implemented a general expression syntax based on the Arden Syntax that has been successfully used to encode eligibility criteria for an application aimed at identifying clinical trials for a patient. Preliminary results from use of our application are encouraging.

References

[1] E-1460 Standard Specification for Defining And Sharing Modular Health Knowledge Bases (Arden Syntax for Medical Logic Modules). ASTM Standards, v 14.01. Philadelphia: ASTM, 1992; 539-87.

[2] Ohno-Machado L; Wang SJ; Mar P; Boxwala A; Decision Support for Clinical Trial Eligibility Determination in Breast Cancer, *submitted, AMIA Fall Symposium 1999.*

[3] Ohno-Machado L, et al; The GuideLine Interchange Format: A Model for Representing Guidelines, *JAMIA,* 1998;5:357-372.

Appendix 3 - Poster

(Poster presented at the 1999 Breast Cancer Research Symposium of the Massachusetts Department of Public Health)

Selection of Clinical Trials Using Artificial Intelligence

INTRODUCTION

Historically, accrual of patients for clinical trials has not been very successful, particularly for certain clinical domains. Studies demonstrate that just a small percentage of eligible patients (3 to 10%) are actually enrolled in such trials. The low accrual rates are attributed to: (1) physician factors such as lack of knowledge about clinical trials, (2) patient factors such as lack of patient-oriented information regarding trials, (3) organizational barriers, and (4) health care system obstacles. If clinical trial information can be made more accessible to patients and their primary care providers (PCPs), we believe that clinical trial accrual rates can improve.

We have developed a WWW-based system for clinical trial eligibility determination where patients or primary care providers can enter clinical information about a patient and obtain a ranked list of clinical trials for which the patient is likely to be eligible.

MATERIAL AND METHODS

Data. We used the National Cancer Institute's Physician Data Query (PDQ) database as the source of information for clinical trials. They contain free-text lists of eligibility criteria organized by patient characteristics (e.g., age, menopausal status); disease characteristics (e.g., histology, metastases); and prior and concurrent therapy. For the preliminary phase of this study, we selected from the PDQ database all Phase II and Phase III trials for the treatment of metastatic or recurrent breast cancer. We chose advanced stage cancer because we hypothesized that these patients would be more interested in seeking participation in clinical trials after exhausting traditional treatment venues. We decided to limit our initial set to Phase II and Phase III trials since these studies are further developed, and typically involve several patients. We found a total of 85 clinical trials in the PDQ database (as of July 1998) that fit these parameters.

Clinical Trial Eligibility Database. Each clinical trial summary was encoded into a structured format. The encoded summary was stored in an XML document. This document contains elements describing identifying information about the clinical trial (name of trial, protocol number) and a collection of criteria elements (Figure 1). Each criterion element contains the original narrative text description from PDQ and the criterion encoded in a computable expression. The criterion is encoded in a modified version of the grammar used for specifying logic statements in the Arden Syntax.

The translation of the original free-text criterion descriptions from PDQ into a machine-interpretable representation was largely a manual process performed by informatics fellows and faculty in our laboratory. We established a uniform basis for encoding criteria.

EXAMPLE

Figures 3 to 6 show an example of the application for a sample breast cancer patient: a premenopausal, 55 year-old woman with stage IV breast cancer with metastases to liver and bone, previous mastectomy, chemotherapy and radiotherapy. This patient also suffers from coronary artery disease and diabetes mellitus. Figure 3 shows the initial data input form in which the PCP has entered some clinical information about the patient. Using this information, the program returns a preliminary list of trials. This list is ranked, with the most likely matches at the top (Figure 4).

If the list is long, the application offers the physician an opportunity to fill in additional patient information to narrow the search. The program dynamically constructs the secondary input form to request the information that would be more likely to narrow the number of clinical trials (Figure 5).

Again, the physician fills in as much additional information as he or she can. This process can be repeated as many times as desired until either the resulting list is short enough, or there is no additional information required or available.

The final list is presented in order of likelihood of match. In this example, the system narrowed the list to 15 trials that the patient is potentially eligible for. A summary of all the entered information is provided. Detailed information about these clinical trials (Figure 6) can be displayed, along with a list of the criteria still to be checked.

Currently, we have not taken into account patient preferences in ranking the clinical trials, such as modality of treatment, potential toxicity, potential for cure, and geographic constraints. The system currently ranks trials solely based on the likelihood that the patient will satisfy the eligibility requirements.

CONCLUSIONS

We have developed a WWW-based decision support system to help patients and providers determine the patient's eligibility for certain clinical trials. The system currently contains all Phase II and III treatment clinical trials for metastatic breast cancer from the NCI's PDQ database. It rules out trials that the patients are not eligible for and ranks the remaining trials according to how many criteria still need to be checked to determine eligibility. This initial prototype system has helped us identify relevant issues in machine-readable criteria representation, user interface design, and clinical trial ranking under uncertainty. Preliminary testing of the system with a few clinical cases has been promising. A formal evaluation of usability and reliability is underway. Future versions of this application will include a belief network that will allow the system to impute missing data values and reason under conditions of uncertainty.

```
<PROTOCOL ID="09251">

<NAME> Phase II Randomized Study of
Cyclophosphamide/Methotrexate/Fluorouracil (CMF)
vs Mitoxantrone in Elderly Patients with Advanced Breast
Cancer
 </NAME>

<!--Disease Characteristics-->
<CRITERION>
 No CNS metastases
 <SPEC>
  (metastases_locations where it is a "CNS") == []
 </SPEC>
</CRITERION>

<!--Patient Characteristics-->
<CRITERION>
Over 70
 <SPEC>
  age &gt; 70
 </SPEC>
</CRITERION>

<CRITERION>
Postmenopausal
  <SPEC>
   menopausal_status == "postmenopausal"
  </SPEC>
</CRITERION>

<!--Hematopoietic:-->
<CRITERION>
WBC at least 3,000
  <SPEC>
  WBC &gt;= 3000
  </SPEC>
</CRITERION>
```

Figure 1. Excerpt of clinical trial protocol structured in XML format.

```
<!-- Patient Characteristics -->

<VARIABLE NAME='age' TYPE='number' CUI='C0001779'>
</VARIABLE>

<VARIABLE NAME='birthdate' TYPE='date' CUI='C0421451'>
</VARIABLE>

<VARIABLE NAME='gender' TYPE='enum' CUI='C0079399'>
Gender of patient
 <VALUE CUI='C0024554'>male</VALUE>
 <VALUE CUI='C0015780'>female</VALUE>
</VARIABLE>

<VARIABLE NAME='menopausal_status' TYPE='enum'
CUI='C0025320'>
Menopausal status of patient.
 <VALUE CUI='C0279752'>premenopausal</VALUE>
 <VALUE CUI='C0279753'>postmenopausal</VALUE>
</VARIABLE>
```

Figure 2. Excerpts from data dictionary containing definitions of clinical concepts used in the eligibility criteria.

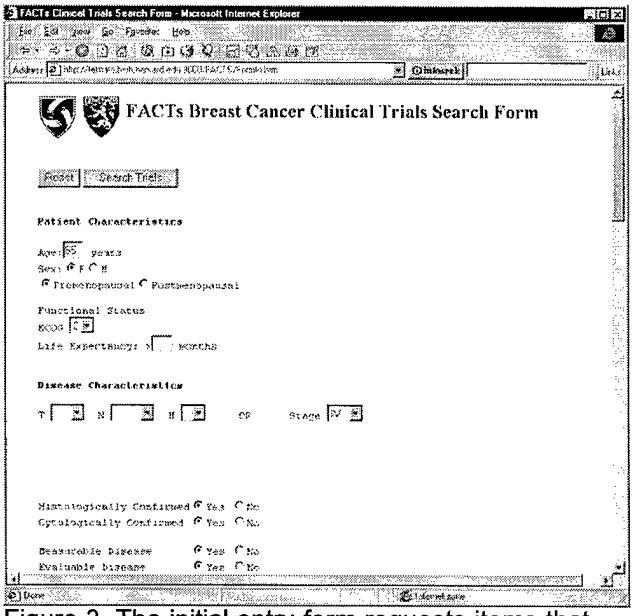


Figure 3. The initial entry form requests items that are most frequent and easiest to obtain.

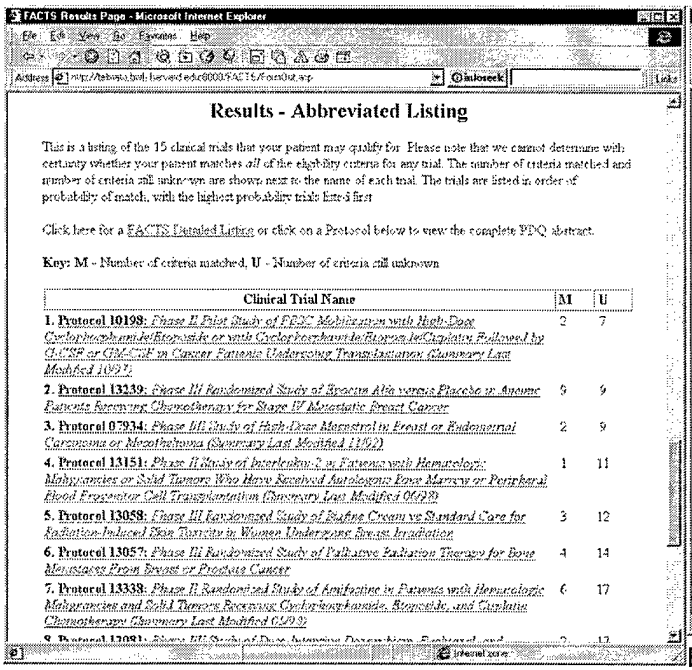


Figure 4. Results page showing a ranked list of clinical trials.

Figure 6. Detailed information about remaining trials is displayed.

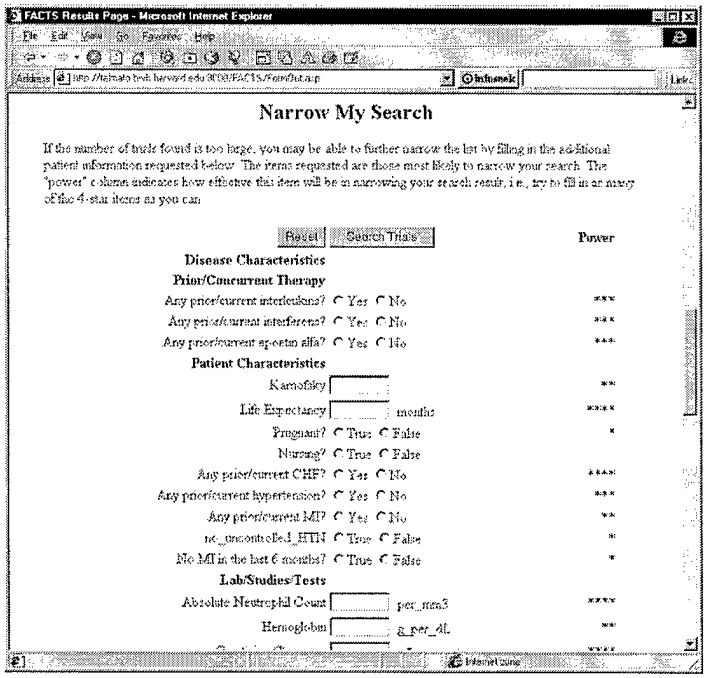


Figure 5. Secondary entry forms are created dynamically and request information that will be most useful in narrowing the search.

Appendix 4 - DTD Protocols

External DTD for FACTs Protocol Files

```
<?xml version="1.0" standalone="yes" encoding="UTF-8"?>
<!--   This is the external DTD for FACTs Protocol Files   -->

<!DOCTYPE FACTS [
<!--  FACTS is the root element of a protocol file, which can  -->
<!--  contain one or more protocols.                           -->

<!ELEMENT  FACTS (PROTOCOL+)>
<!--  The root can contain one or more PROTOCOL Elements,  -->
<!--  each of which  contains a single protocol.           -->

<!ELEMENT PROTOCOL (NAME, CRITERION+)>
<!--  The PROTOCOL element contains a NAME (of the NCI        -->
<!--  protocol) and one or more CRITERION elements (criteria  -->
<!--  for entry into the NCI protocol).                        -->

<!ATTLIST PROTOCOL
      ID ID #REQUIRED>
<!--  The PROTOCOL element has an ID (the NCI protocol number)  -->
<!--  attribute, which is unique and required.                   -->

<!ELEMENT NAME (#PCDATA)>
<!--  The NAME (as listed in the NCI protocol) element consists of  -->
<!--  parsed character data.                                        -->

<!ELEMENT CRITERION (#PCDATA, SPEC)>
<!--  The CRITERION element contains parsed character data        -->
<!--  (a description of the criterion) and a SPEC element (which  -->
<!--  delimits the specification variable which will pass to the  -->
<!--  parser).                                                    -->

<!ELEMENT SPEC (#PCDATA)>
<!--  The SPEC element contains parsed character data, consisting  -->
<!--  of a specification variable, which will pass to the parser.  -->

]>
```

Appendix 5 - DTD Variables

External DTD for FACTs Variable Files

```
<?xml version="1.0" standalone="yes" encoding="UTF-8"?>
<!--   This is the external DTD for FACTs Variable Files   -->

<!DOCTYPE FACTS [
<!--  FACTS is the root element of the variable file, which  -->
<!--  can contain Type Declarations, Variables, and Fields.  -->

<!ENTITY choices "CDATA | 'struct' | 'enum' | 'list' | 'number' | 'string' |
'date' | 'duration'">
<!--  The 'choices' entity lists the choices for the string type  -->
<!--  of the TYPE and ELEMENT-TYPE attributes.                    -->

<!ELEMENT  FACTS (TYPE-DECL | VARIABLE | FIELD)*)>
<!--  The root can contain zero or more TYPE-DECL, VARIABLE,  -->
<!--  and/or FIELD elements in any order.                     -->

<!ELEMENT TYPE-DECL (#PCDATA, (VALUE* | FIELD*))>
<!--  The TYPE-DECL element contains a description of the Type   -->
<!--  Declaration, consisting of parsed character data, and zero -->
<!--  or more of either VALUE or FIELD elements.                 -->

<!ATTLIST TYPE-DECL
      NAME CDATA #REQUIRED
      TYPE (&choices;) #REQUIRED
      UNITS CDATA #IMPLIED
      ELEMENT-TYPE (&choices;) #IMPLIED>
<!--  The TYPE-DECL element has four attributes:                        -->
<!--       1.  NAME, consisting of character data, which is required  -->
<!--           and has no default.                                    -->
<!--       2.  TYPE, which consists of one of the strings defined by  -->
<!--           the 'choices' entity, which is required and has no     -->
<!--           default value.                                         -->
<!--       3.  UNITS, which consists of character data, and which is  -->
<!--           optional and without a default value.                  -->
<!--       4.  ELEMENT-TYPE, which consists of one of the strings     -->
<!--           defined by the 'choices' entity, which is optional     -->
<!--           and has no default value.                              -->

<!ELEMENT VARIABLE (#PCDATA, (VALUE* | FIELD*))>
<!--  The VARIABLE element contains a description of the Variable,   -->
<!--  consisting of parsed character data, and zero or more of       -->
<!--  either VALUE or FIELD elements.                                -->

<!ATTLIST VARIABLE
      NAME CDATA #REQUIRED
      TYPE (&choices;) #REQUIRED
      UNITS CDATA #IMPLIED
      ELEMENT-TYPE (&choices;) #IMPLIED>
<!--  The VARIABLE element has four attributes:                         -->
<!--       1.  NAME, consisting of character data, which is required  -->
```

```
<!--            and has no default.                                    -->
<!--        2.  TYPE, which consists of one of the strings defined by  -->
<!--            the 'choices' entity, which is required and has no     -->
<!--            default value.                                         -->
<!--        3.  UNITS, which consists of character data, and which is  -->
<!--            optional and without a default value.                  -->
<!--        4.  ELEMENT-TYPE, which consists of one of the strings     -->
<!--            defined by the 'choices' entity, which is optional     -->
<!--            and has no default value.                              -->

<!ELEMENT FIELD (#PCDATA, (VALUE* | FIELD*))>
<!--  The FIELD element contains a description of the Field,     -->
<!--  consisting of parsed character data, and zero or more of  -->
<!--  either VALUE or FIELD elements.                           -->

<!ATTLIST FIELD
      NAME CDATA #REQUIRED
      TYPE (&choices;) #REQUIRED
      UNITS CDATA #IMPLIED
      ELEMENT-TYPE (&choices;) #IMPLIED>
<!--  The FIELD element has four attributes:                          -->
<!--        1.  NAME, consisting of character data, which is required -->
<!--            and has no default.                                   -->
<!--        2.  TYPE, which consists of one of the strings defined by -->
<!--            the 'choices' entity, which is required and has no    -->
<!--            default value.                                        -->
<!--        3.  UNITS, which consists of character data, and which is -->
<!--            optional and without a default value.                 -->
<!--        4.  ELEMENT-TYPE, which consists of one of the strings    -->
<!--            defined by the 'choices' entity, which is optional    -->
<!--            and has no default value.                             -->

<!ELEMENT VALUE (#PCDATA, IS-A*)>
<!--  The VALUE element contains a description of the value, consisting  -->
<!--  of parsed character data, and zero or more IS-A elements.          -->

<!ELEMENT IS_A (#PCDATA)>
<!--  The IS-A element is composed of parsed character data.  -->

]>
```

Appendix 6 - Example of Encoded Protocol

Example of Protocol File Encoded in XML

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!--  This is the internal DTD for FACTs Protocol Files  -->
<!DOCTYPE FACTS SYSTEM "FACTsProtocolExternalDTD.dtd">


<FACTS>

<PROTOCOL ID="11092">

  <NAME>
Phase II Study of Sequential High-Dose Cyclophosphamide, Melphalan, and
Thiotepa Followed by Peripheral Blood Stem Cell Rescue in
Chemotherapy-Sensitive Metastatic Breast Cancer (Summary Last Modified 07/97)
  </NAME>

<!--Disease Characteristics-->


  <CRITERION>
 Histologically documented breast carcinoma that is stage IV and considered
incurable by standard treatment
  Pathology reviewed by Yale-New Haven Hospital Department of Pathology
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>



  <CRITERION>
 Ongoing objective response to prior induction chemotherapy required
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>



  <CRITERION>
 No brain metastasis
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>


<!--Hormone receptor status:-->

  <CRITERION>
 Estrogen and progesterone status known
```

```
        <SPEC>
           unknown
        </SPEC>
    </CRITERION>


<!--Prior/Concurrent Therapy-->

<!--Biologic therapy:-->
    <!--Not specified-->

<!--Chemotherapy:-->

    <CRITERION>
 No prior mitomycin or nitrosourea
        <SPEC>
           unknown
        </SPEC>
    </CRITERION>


    <CRITERION>
 At least 4 weeks since chemotherapy at time of stem cell harvest
        <SPEC>
           unknown
        </SPEC>
    </CRITERION>


<!--Endocrine therapy:-->
    <!--Not specified-->

<!--Radiotherapy:-->

    <CRITERION>
 No prior radiotherapy to pelvis or brain
        <SPEC>
           unknown
        </SPEC>
    </CRITERION>


<!--Surgery:-->

    <CRITERION>
 At least 2 weeks since major surgery
        <SPEC>
           unknown
        </SPEC>
    </CRITERION>


<!--Patient Characteristics-->

<!--Age:-->

    <CRITERION>
```

```
Over 18
      <SPEC>
       age &gt;  18
      </SPEC>
  </CRITERION>


<!--Sex:-->

  <CRITERION>
Women or men
      <SPEC>
         unknown
      </SPEC>
  </CRITERION>


<!--Menopausal status:-->
  <!--Not specified-->

<!--Performance status:-->

  <CRITERION>
Karnofsky 80%-100%
      <SPEC>
         unknown
      </SPEC>
  </CRITERION>


<!--Life expectancy:-->

  <CRITERION>
Greater than 3 months
      <SPEC>
         unknown
      </SPEC>
  </CRITERION>


<!--Hematopoietic:-->

  <CRITERION>
ANC greater than 1,500
      <SPEC>
           ANC &gt;= 1500
      </SPEC>
  </CRITERION>


  <CRITERION>
Platelets greater than 100,000
      <SPEC>
         unknown
      </SPEC>
  </CRITERION>
```

```
  <CRITERION>
 Hemoglobin greater than 9 g/dL
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>


<!--Hepatic:-->

  <CRITERION>
 Bilirubin less than 1.5 times normal (unless benign congenital
     hyperbilirubinemia)
       <SPEC>
      if bilirubin &lt;= 1.5 x_nl_bilirubin then
       conclude true;
      endif;
       </SPEC>
  </CRITERION>


  <CRITERION>
 PT and PTT normal
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>


  <CRITERION>
 Liver biopsy normal if serologic evidence of active hepatitis B or C
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>


<!--Renal:-->

  <CRITERION>
 Creatinine no greater than 1.2 mg/dL
       <SPEC>
       creatinine &lt;= 1.2 mg_per_dL
       </SPEC>
  </CRITERION>


<!--Cardiovascular:-->

  <CRITERION>
 No abnormal wall motion
       <SPEC>
          unknown
       </SPEC>
  </CRITERION>
```

```
 <CRITERION>
No active heart disease
      <SPEC>
         unknown
      </SPEC>
 </CRITERION>


 <CRITERION>
Left ventricular ejection fraction at least 50%
      <SPEC>
         unknown
      </SPEC>
 </CRITERION>


<!--Pulmonary:-->

 <CRITERION>
DLCO normal
      <SPEC>
         unknown
      </SPEC>
 </CRITERION>


<!--Other:-->

 <CRITERION>
Nutritional status adequate (greater than 1,000 calories/day orally)
      <SPEC>
         unknown
      </SPEC>
 </CRITERION>


 <CRITERION>
No HIV infection
      <SPEC>
           not HIV
      </SPEC>
 </CRITERION>


 <CRITERION>
No other active serious medical or psychiatric disease
      <SPEC>
         unknown
      </SPEC>
 </CRITERION>


 <CRITERION>
No other malignancy except:
    Basal cell skin carcinoma
    In situ cervical cancer
```

```
      <SPEC>
         unknown
      </SPEC>
  </CRITERION>


  <CRITERION>
 Negative pregnancy test required of fertile women
      <SPEC>
       not pregnant
      </SPEC>
  </CRITERION>


  <CRITERION>
 Barrier method birth control required of fertile women throughout study and
     for up to 2 years thereafter
      <SPEC>
         unknown
      </SPEC>
  </CRITERION>



</PROTOCOL>

</FACTS>
```

Appendix 7 - Variables encoded

Variables encoded in XML files

```
<?xml version="1.0" encoding="UTF-8"
standalone="no"?>
<!DOCTYPE FACTS SYSTEM
"FACTsVariableExternalDTD.dtd">
<FACTS>


<!-- Disease Characteristics -->

<!-- cancer type, how diagnosed,
size, location, count -->
<VARIABLE NAME='breast_cancer_type'
TYPE='list' ELEMENT-TYPE='enum'>
Type of breast cancer.
  <VALUE>DCIS</VALUE>
  <VALUE>LCIS</VALUE>
  <VALUE>mixed DCIS and LCIS</VALUE>
  <VALUE>adenocarcinoma</VALUE>
  <VALUE>squamous cell
carcinoma</VALUE>
  <VALUE>apocrine</VALUE>
  <VALUE>adenoidcystic</VALUE>
</VARIABLE>

<VARIABLE NAME='diagnosis'
TYPE='string'>
Dummy variable to hold diagnosis
date.
Use the function
start_date(diagnosis) to get the
diagnosis date.
Date of initial diagnosis of breast
cancer.
</VARIABLE>

<!--  SJW resurrected from graveyard
-->
<VARIABLE
NAME="histologically_confirmed"
TYPE="boolean">
  <DISPLAY>Breast CA confirmed by
histology?</DISPLAY>
Histology result positive
</VARIABLE>

<VARIABLE
NAME="cytologically_confirmed"
TYPE="boolean">
  <DISPLAY>Breast CA confirmed by
cytology?</DISPLAY>
confirmed breast CA by cytology
</VARIABLE>

<!-- removed by SJW
<VARIABLE NAME='diagnosis_method'
TYPE='list' ELEMENT-TYPE='enum'>
Procedures used to diagnose primary
tumor.
  <VALUE>biopsy of metastatic
site</VALUE>
  <VALUE>biopsy of primary
site</VALUE>
  <VALUE>CT scan</VALUE>
  <VALUE>core needle biopsy</VALUE>
  <VALUE>histology</VALUE>
  <VALUE>FNA</VALUE>
  <VALUE>cytology</VALUE>
  <VALUE>immunohistochemistry</VALUE>
  <VALUE>mammography</VALUE>
  <VALUE>MRI</VALUE>
  <VALUE>physical exam</VALUE>
  <VALUE>ultrasound</VALUE>
  <VALUE>CXR</VALUE>
</VARIABLE>
-->

<!-- current status -->

<!-- Note that cancer_status also
will have a "since" date specified. -
->
<!-- SJW removed cancer_status
<!-- But do we need it for
start_date?
<VARIABLE NAME='cancer_status'
TYPE='enum' ELEMENT-TYPE='enum'>
Current state of primary breast
cancer.
</VARIABLE>
-->
<!-- SJW removed
  <VALUE>disease free</VALUE>   or
remission
  <VALUE>measurable disease</VALUE>
  <VALUE>evaluable disease</VALUE>
but not measurable
  <VALUE>recurrent</VALUE>
  <VALUE>progressing</VALUE>
  <VALUE>rapidly progressing
   <IS-A>progressing</IS-A>
```

```
  </VALUE>
-->

<VARIABLE NAME='measurable_disease'
TYPE='boolean'>
  <DISPLAY>Measurable
Disease?</DISPLAY>
Must be bidimensionally measurable.
The following do not count as
measurable:
  Bone marrow blastic lesions
  pleural effusions
  previously irradiated lesions
  histologically verified only
  lymphangitic disease

</VARIABLE>

<VARIABLE NAME='evaluable_disease'
TYPE='boolean'>
  <DISPLAY>Evaluable
Disease?</DISPLAY>
Measurable disease is always
evaluable.
</VARIABLE>

<VARIABLE NAME='NED' TYPE='boolean'>
  <DISPLAY>No Evidence of
Disease?</DISPLAY>
No Evidence of Disease
(equivalent to disease_free?)
</VARIABLE>

<VARIABLE NAME='recurrent'
TYPE='boolean'>
  <DISPLAY>Is the cancer
recurrent?</DISPLAY>
</VARIABLE>

<VARIABLE NAME='progressing'
TYPE='boolean'>  <!-- includes
rapidly progressing? -->
  <DISPLAY>Progressing?</DISPLAY>
</VARIABLE>

<VARIABLE NAME='rapidly_progressing'
TYPE='boolean'>  <!-- includes
rapidly progressing? -->
</VARIABLE>

<VARIABLE NAME='locally_advanced'
TYPE='boolean'>
Includes  T3, T4,
 fixed ipsilateral axillary nodes,
 supra or infraclavicular nodes,
 arm edema
</VARIABLE>

<VARIABLE NAME='resectable'
TYPE='boolean'>
(SJW: Same as 'operable'?)
</VARIABLE>

<VARIABLE NAME='operable'
TYPE='boolean'>
Defined as Stage IIIa (AJC 1983) or
below.
</VARIABLE>

<VARIABLE NAME='curable'
TYPE='boolean'>
definition??
</VARIABLE>

<VARIABLE NAME='tumor_size'
TYPE='number' UNITS='cm'>
</VARIABLE>


<!-- Staging -->

<!--SJW changed from number to enum
and incorporated subtype -->
<VARIABLE NAME='Stage' TYPE='enum'>
Clinical stage
  <DISPLAY>Stage</DISPLAY>
  <VALUE>1</VALUE>
  <VALUE>2</VALUE>
  <VALUE>2A
   <IS-A>2</IS-A>
  </VALUE>
  <VALUE>2B
   <IS-A>2</IS-A>
  </VALUE>

  <VALUE>3</VALUE>
  <VALUE>3A
   <IS-A>3</IS-A>
  </VALUE>
  <VALUE>3B
   <IS-A>3</IS-A>
  </VALUE>
  <VALUE>4</VALUE>
</VARIABLE>


<!-- SJW removed
<VARIABLE NAME='Stage_subtype'
TYPE='enum'>
Clinical stage subtype
  <VALUE>a</VALUE>
  <VALUE>b</VALUE>
</VARIABLE>
-->
```

```
<!--SJW changed from number to enum
and incorporated subtype -->
<VARIABLE NAME='T' TYPE='enum'>
T-stage of the AJCC's TNM staging
system.
<!-- Do we need a TX? -->
  <VALUE>0</VALUE>
  <VALUE>is</VALUE>
  <VALUE>1</VALUE>
  <VALUE>1a
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1b
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1c
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>2</VALUE>
  <VALUE>3</VALUE>
  <VALUE>4</VALUE>
  <VALUE>4a
   <IS-A>4</IS-A>
  </VALUE>
  <VALUE>4b
   <IS-A>4</IS-A>
  </VALUE>
  <VALUE>4c
   <IS-A>4</IS-A>
  </VALUE>
  <VALUE>4d
   <IS-A>4</IS-A>
  </VALUE>
</VARIABLE>


<!--
<VARIABLE NAME='T_subtype'
TYPE='enum'>
T subtype.

Some combinations are invalid.
   Valid combinations:
   T1a,T1b,T1c, T4a,T4b,T4c,T4d
   To code "Tis" use T=0,
T_subtype="is"

  <VALUE>is</VALUE>
  <VALUE>a</VALUE>
  <VALUE>b</VALUE>
  <VALUE>c</VALUE>
  <VALUE>d</VALUE>
</VARIABLE>
-->
```

```
<!--SJW changed from number to enum
and incorporated subtype -->

<VARIABLE NAME='N' TYPE='enum'>
N-stage of the AJCC's TNM staging
system.
  <VALUE>0</VALUE>
  <VALUE>1</VALUE>
  <VALUE>1a
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1b
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1bi
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1bii
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1biii
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>1biv
   <IS-A>1</IS-A>
  </VALUE>
  <VALUE>2</VALUE>
</VARIABLE>


<!-- N stages can be either clinical
or pathological.
Do we need a separate "pN" stage or
can we do it like this? -->


<!--
<VARIABLE NAME='N_subtype'
TYPE='enum'>
N-subtype for pathological staging.
Note that these only apply for pN1.
  <VALUE>a</VALUE>
  <VALUE>b</VALUE>
  <VALUE>bi</VALUE>
  <VALUE>bii</VALUE>
  <VALUE>biii</VALUE>
  <VALUE>biv</VALUE>
</VARIABLE>
-->

<!--SJW changed from number to enum
and incorporated subtype -->
<VARIABLE NAME='M' TYPE='number'>
M-stage of the AJCC's TNM staging
system.
  <VALUE>0</VALUE>
  <VALUE>1</VALUE>
```

```
</VARIABLE>

<VARIABLE NAME='grade' TYPE='number'>
Histological Grade of primary tumor.
      G1 well-differentiated (low
grade),
      G2 moderately differentiated,
      G3 poorly differentiated,
      G4 undifferentiated (high
grade)
  <VALUE>1</VALUE>
  <VALUE>2</VALUE>
  <VALUE>3</VALUE>
  <VALUE>4</VALUE>
</VARIABLE>

<VARIABLE NAME='tumor_count'
TYPE='number'>
Number of tumors
</VARIABLE>

<VARIABLE NAME='margins' TYPE='enum'>
Margins
  <VALUE>positive</VALUE>
  <VALUE>negative</VALUE>
</VARIABLE>


<!-- -Lymph Nodes -->

<VARIABLE
NAME='positive_axillary_node_count'
TYPE='number'>
Number of positive axillary lymph
nodes.
</VARIABLE>

<VARIABLE
NAME='total_axillary_node_count'
TYPE='number'>
Total number of axillary lymph nodes
sampled/dissected.
</VARIABLE>

<!-- These two should specify
"internal_mammary_..."  -->
<!-- Are these two really needed?? --
>
<VARIABLE
NAME='positive_internal_mammary_node_
count' TYPE='number'>
Number of positive internal mammary
lymph nodes
</VARIABLE>

<VARIABLE
NAME='total_internal_mammary_node_cou
nt' TYPE='number'>
Total number of internal mammary
lymph nodes dissected
</VARIABLE>



<!-- Metastases -->

<!-- whether or not metastases are
present
     Is this exactly the same as M
stage?   -->
<VARIABLE NAME='metastases_present'
TYPE='boolean'>
  <DISPLAY>Metastases
present?</DISPLAY>
</VARIABLE>

<VARIABLE NAME='metastases_locations'
TYPE='list' ELEMENT-TYPE='enum'>
  <DISPLAY>Metastases
Locations:</DISPLAY>


Locations of distant metastases.
  <VALUE>bone</VALUE>
  <VALUE>bone marrow
   <IS-A>bone</IS-A>
  </VALUE>
  <VALUE>lytic bone lesions
   <IS-A>bone</IS-A>
  </VALUE>
  <VALUE>blastic bone lesions
   <IS-A>bone</IS-A>
  </VALUE>
  <VALUE>CNS</VALUE>
  <VALUE>brain
   <IS-A>CNS</IS-A>
  </VALUE>
  <VALUE>meninges
   <IS-A>CNS</IS-A>
  </VALUE>
  <VALUE>viscera</VALUE>
  <VALUE>liver
   <IS-A>viscera</IS-A>
  </VALUE>
  <VALUE>lung
   <IS-A>viscera</IS-A>
  </VALUE>

  <VALUE>distant lymph nodes</VALUE>
<!-- does not refer to nodes in node
count section -->
```

```
  <VALUE>ipsilateral distant lymph
nodes <!-- includes supraclavicular
and cervical -->
   <IS-A>distant lymph nodes</IS-A>
  </VALUE>
  <VALUE>internal mammary lymph nodes
   <IS-A>ipsilateral distant lymph
nodes</IS-A>
  </VALUE>
</VARIABLE>


<VARIABLE NAME='other_findings'
TYPE='list' ELEMENT-TYPE='enum'>
  <VALUE>fixed to skin</VALUE>
  <VALUE>fixed to chest wall</VALUE>
  <VALUE>edema</VALUE>
  <VALUE>peau d_orange</VALUE>
  <VALUE>ulceration</VALUE>
  <VALUE>erythema</VALUE>
  <VALUE>dimpling</VALUE>
  <VALUE>nipple inversion or
retraction</VALUE>
  <VALUE>pagets disease</VALUE>
  <VALUE>inflammatory</VALUE>
  <VALUE>poor-risk features</VALUE>
</VARIABLE>


<!-- Hormone receptor status -->
<VARIABLE
NAME='progesterone_receptor'
TYPE='enum'>
  <DISPLAY>Progesterone Receptor
status</DISPLAY>

  <VALUE>positive</VALUE>
  <VALUE>negative</VALUE>
</VARIABLE>

<VARIABLE NAME='estrogen_receptor'
TYPE='enum'>
  <DISPLAY>Estrogen Receptor
status</DISPLAY>
  <VALUE>positive</VALUE>
  <VALUE>negative</VALUE>
</VARIABLE>

<VARIABLE NAME='HER2_neu_receptor'
TYPE='enum'>
  <VALUE>positive</VALUE>
  <VALUE>negative</VALUE>
</VARIABLE>

<VARIABLE NAME='hormone_resistant'
TYPE='boolean'>
Whether the breast CA has been
determined to be hormone resistant.
</VARIABLE>

<!-- Prior/Concurrent Therapy -->

<VARIABLE NAME='treatments'
TYPE='list' ELEMENT-TYPE='struct'>
This contains all of the patient's
prior and concurrent treatments,
therapies, or medications.

 <FIELD NAME='name' TYPE='enum'>
  <VALUE>biological therapy</VALUE>

  <VALUE>colony stimulating factor
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>interleukins
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>interferons
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>bone marrow transplant
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>stem cell transplant
   <IS-A>biological therapy</IS-A>
  </VALUE>

  <VALUE>radiotherapy</VALUE>  <!--
XRT -->

  <VALUE>radiation bolus
   <IS-A>radiotherapy</IS-A>
  </VALUE>
  <VALUE>brachytherapy boost
   <IS-A>radiotherapy</IS-A>
  </VALUE>


  <VALUE>chemotherapy</VALUE>
  <VALUE>cytoxan
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>CAF
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>CMF
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>cyclophosphamide
  <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>mitomycin
  <IS-A>chemotherapy</IS-A>
```

```
  </VALUE>
  <VALUE>platinum-based chemotherapy
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>dose-intensification
regimens
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>bisphosphonate
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>nitrosoureas
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>carmustine
   <IS-A>chemotherapy</IS-A>
  </VALUE>



  <VALUE>anthraquinone     <!-- need
to find out which drugs are
anthraquinones -->
   <IS-A>chemotherapy</IS-A>
  </VALUE>

  <VALUE>anthracine     <!-- need to
find out which drugs are anthracines
-->
   <IS-A>chemotherapy</IS-A>
  </VALUE>

  <VALUE>anthracycline
   <IS-A>chemotherapy</IS-A>
  </VALUE>
  <VALUE>doxorubicin
   <IS-A>anthracycline</IS-A>
  </VALUE>
  <VALUE>mitoxantrone
   <IS-A>anthracycline</IS-A>
  </VALUE>
<VALUE>epirubicin
   <IS-A>anthracycline</IS-A>
  </VALUE>
  <VALUE>pirarubicin
   <IS-A>anthracycline</IS-A>
  </VALUE>



  <VALUE>endocrine therapy</VALUE>
  <VALUE>HRT
   <IS-A>endocrine therapy</IS-A>
  </VALUE>

  <VALUE>contraceptive</VALUE>
  <VALUE>oral contraceptives
   <IS-A>contraceptive</IS-A>
  </VALUE>

  <VALUE>androgens
   <IS-A>endocrine therapy</IS-A>
  </VALUE>
  <VALUE>estrogens
   <IS-A>endocrine therapy</IS-A>
  </VALUE>
  <VALUE>phytoestrogens
   <IS-A>estrogens</IS-A>
  </VALUE>
  <VALUE>antiestrogens
   <IS-A>endocrine therapy</IS-A>
  </VALUE>
   <VALUE>Tamoxifen
    <IS-A>antiestrogens</IS-A>
  </VALUE>
  <VALUE>corticosteroids
   <IS-A>endocrine therapy</IS-A>
  </VALUE>
  <VALUE>progestational agents
   <IS-A>endocrine therapy</IS-A>
  </VALUE>
  <VALUE>megestrol acetate
   <IS-A>endocrine therapy</IS-A>
  </VALUE>

  <VALUE>surgery</VALUE>
  <VALUE>tumorectomy
   <IS-A>lumpectomy</IS-A>
  </VALUE>
  <VALUE>lumpectomy
   <IS-A>surgery</IS-A>
  </VALUE>
  <!--lumpectomy with level 1
axillary dissection-->
  <VALUE>lumpectomy L1
   <IS-A>lumpectomy</IS-A>
  </VALUE>
  <!--lumpectomy with level 1 and 2
axillary dissection-->
  <VALUE>lumpectomy L2
   <IS-A>lumpectomy</IS-A>
  </VALUE>
  <!--lumpectomy with level 1, 2 and
3 axillary dissection (axillary
clearance) -->
  <VALUE>lumpectomy L3
   <IS-A>lumpectomy</IS-A>
  </VALUE>
  <VALUE>segmental mastectomy
   <IS-A>lumpectomy</IS-A>
   <IS-A>mastectomy</IS-A>
  </VALUE>
  <VALUE>quadrantectomy
   <IS-A>lumpectomy</IS-A>
  </VALUE>
```

```
 <VALUE>mastectomy
  <IS-A>surgery</IS-A>
 </VALUE>
 <VALUE>modified radical mastectomy
  <IS-A>mastectomy</IS-A>
 </VALUE>
 <VALUE>radical mastectomy
  <IS-A>mastectomy</IS-A>
 </VALUE>
 <VALUE>total mastectomy
  <IS-A>mastectomy</IS-A>
 </VALUE>

 <VALUE>Hysterectomy
  <IS-A>surgery</IS-A>
 </VALUE>
 <VALUE>oophorectomy
  <IS-A>surgery</IS-A>
 </VALUE>

 <VALUE>adrenalectomy
  <IS-A>surgery</IS-A>
 </VALUE>
 <VALUE>hypophysectomy
  <IS-A>surgery</IS-A>
 </VALUE>

 <VALUE>epoetin alfa
 </VALUE>
 <VALUE>aromatase inhibitor
 </VALUE>
  <VALUE>Faslodex
   <IS-A>aromatase inhibitor</IS-A>
 </VALUE>
  <VALUE>anastrozole
   <IS-A>aromatase inhibitor</IS-A>
 </VALUE>
 <VALUE>investigational drug NOS
 </VALUE>

 <VALUE>anticoagulant</VALUE>

</FIELD>


<FIELD NAME='type' TYPE='enum'>
 <VALUE>primary</VALUE>
 <VALUE>adjuvant</VALUE>
 <VALUE>neoadjuvant</VALUE>
 <VALUE>palliative</VALUE>
 <VALUE>systemic</VALUE>
 <VALUE>topical</VALUE>

</FIELD>

<FIELD NAME='body_site' TYPE='enum'>
  <VALUE>locoregional</VALUE>  <!--
eg for radiotherapy -->
  <VALUE>pelvic bones</VALUE>
  <VALUE>spine</VALUE>
  <VALUE>lower spine
   <IS-A>spine</IS-A>
  </VALUE>
  <VALUE>mediastinum</VALUE>
  <VALUE>parasternal nodes</VALUE>
  <VALUE>bone marrow</VALUE>
  <VALUE>bone</VALUE>
</FIELD>

 <FIELD NAME='indication'
TYPE='enum'>
  <VALUE>metastases</VALUE>
  <VALUE>recurrence</VALUE>
  <VALUE>local recurrence
   <IS-A>recurrence</IS-A>
  </VALUE>
 </FIELD>

 <FIELD NAME='number_of_treatments'
TYPE='number'>
 </FIELD>

 <FIELD NAME='frequency' TYPE='enum'>
  <VALUE>once per week</VALUE>
 </FIELD>

 <FIELD NAME='cumulative_dose'
TYPE='number'>
 For chemotherapy.
 </FIELD>

</VARIABLE>



<!-- Patient Characteristics -->

<!-- Age -->
<VARIABLE NAME='Age' TYPE='number'>
</VARIABLE>

<VARIABLE NAME='birthdate'
TYPE='date'></VARIABLE>


<!-- Sex -->
<VARIABLE NAME='gender' TYPE='enum'
WEIGHT = '0.1'>
Gender of patient
  <VALUE>male</VALUE>
  <VALUE>female</VALUE>
</VARIABLE>
```

```
<!-- Menopausal Status -->
<!-- SJW removed perimenopausal -->
<VARIABLE NAME='menopausal_status'
TYPE='enum'>
Menopausal status of patient
  <DISPLAY>Menopausal
Status</DISPLAY>
  <VALUE>premenopausal</VALUE>
  <VALUE>postmenopausal</VALUE>
</VARIABLE>


<!-- Performance Status -->
<VARIABLE NAME='ECOG' TYPE='number'>
<!-- SJW: Is this equivalent to WHO?
-->
Eastern Cooperative Oncology Group
Performance Status

Grade

0     Fully active, able to carry on
all pre-disease performance without
restriction (corresponds to Karnofsky
100)
1     Restricted in physically
strenuous activity but ambulatory and
able to carry out work of a light or
sedentary nature, e.g., light house
work, office work (corresponds to
Karnofsky 90-80)
2     Ambulatory and capable of all
selfcare but unable to carry out any
work activities. Up and about more
than 50% of waking hours (corresponds
to Karnofsky 70-60)
3     Capable of only limited
selfcare, confined to bed or chair
more than 50% of waking hours
(corresponds to Karnofsky 50-40)
4     Completely disabled. Cannot
carry on any selfcare. Totally
confined to bed or chair (corresponds
to Karnofsky 30-20)
5     Dead

As published in Am. J. Clin. Oncol.
(CCT) 5:649-655, 1982
Corresponding data for Karnofsky
added by JJT from

<!-- SJW Summ Formula needs to be
tested -->
    <SUMMARY-FORMULA>
      if Karnofsky == 100 then
        conclude 0;
      endif;
      if ((100 &gt; Karnofsky) and
(Karnofsky &gt;= 80)) then
        conclude 1;
      endif;
      if ((80 &gt; Karnofsky) and
(Karnofsky &gt;= 60)) then
        conclude 2;
      endif;
      if ((60 &gt; Karnofsky) and
(Karnofsky &gt;= 40)) then
        conclude 3;
      endif;
      if ((40 &gt; Karnofsky) and
(Karnofsky &gt;= 20)) then
        conclude 4;
      endif;
      if (20 &gt; Karnofsky) then
        conclude 5;
      endif;
    </SUMMARY-FORMULA>

</VARIABLE>

<!-- Functional Status -->
<VARIABLE NAME='Zubrod'
TYPE='number'>
</VARIABLE>

<!-- Functional Status -->
<VARIABLE NAME='CALGB'
TYPE='number'></VARIABLE>
<VARIABLE NAME='Karnofsky'
TYPE='number'>

<!-- added by JJT, from Karnofsky,
David A. and Burchenal, Joseph H.,
1948 -->

Karnofsky Description Scale (%)

Normal, no complaint 100
Able to carry on normal activities;
minor signs or symptoms of disease 90
Normal activity with effort 80
Cares for self. Unable to carry on
normal activity or to do active work
70
Ambulatory.Requires some assistance
but able to care for most of own
needs 60
Requires considerable assistance and
frequent medical care 50
Disabled; requires special care and
assistance 40
```

```
Severely disabled, hospitalization
indicated though death not imminent
30
Very sick. Hospitalization necessary.
Active supportive treatment necessary
20
Moribund 10
Dead 0
</VARIABLE>

<VARIABLE NAME='prognosis'
TYPE='enum'>
 <VALUE>excellent</VALUE>
 <VALUE>good</VALUE>
 <VALUE>fair</VALUE>
 <VALUE>poor</VALUE>
</VARIABLE>

<!-- Life Expectancy -->
<VARIABLE NAME='life_expectancy'
TYPE='number' UNITS='months'>
  <DISPLAY>Life Expectancy</DISPLAY>
</VARIABLE>


<!-- Substance abuse -->
<VARIABLE NAME='alcoholism'
TYPE='boolean'>
Whether patient is currently abusing
alcohol
</VARIABLE>

<VARIABLE NAME='drug_abuse'
TYPE='boolean'>
Whether patient currently abuses
drugs
</VARIABLE>


<!-- conditions -->
<VARIABLE NAME='pregnant'
TYPE='boolean' WEIGHT='0.1'>
Whether patient is pregnant
  <DISPLAY>Pregnant?</DISPLAY>
</VARIABLE>

<VARIABLE NAME='nursing'
TYPE='boolean' WEIGHT='0.1'>
Whether patient is nursing
  <DISPLAY>Nursing?</DISPLAY>
</VARIABLE>

<VARIABLE
NAME='central_venous_access'
TYPE='boolean'>
Whether patient has a central
catheter
</VARIABLE>

<VARIABLE NAME='fertile'
TYPE='boolean'>
Whether patient is able to conceive.
This is different from premenopausal.
</VARIABLE>

<VARIABLE NAME=
'effective_contraception'
TYPE='boolean'>
Whether patient is using some form of
pregnancy prevention.  This is
different from contraception under
treatments, because it includes
sterilization, barrier methods, and
abstinence.
</VARIABLE>


<!-- contraindications, allergies,
intolerances -->
<VARIABLE
NAME='treatment_intolerances'
TYPE='list' ELEMENT-TYPE='enum'>
List of medications or other
treatments that are contraindicated
for this patient because of allergy,
adverse reaction, immune status, or
any other reason.
  <VALUE>tamoxifen</VALUE>
  <VALUE>mammalian cell derived
product</VALUE>
  <VALUE>eggs</VALUE>
  <VALUE>murine protein</VALUE>
</VARIABLE>



<!--
********************************** -
->
<!-- MEDICAL HISTORY / CURRENT
ILLNESSES -->
<!--
********************************** -
->

<VARIABLE
NAME='serious_medical_condition'
TYPE='boolean'>
Whether any other serious medical
problems exist.
Usually this is in the context of
whether there is a medical condition
```

```
that would preclude this patient from
being a surgical candidate.
  <DISPLAY>Serious medical condition
that would preclude
surgery?</DISPLAY>
</VARIABLE>


<!-- Sam's test variable only -->
<VARIABLE NAME='sodium'
TYPE='struct'>
 <FIELD NAME='result' TYPE='number'>
 </FIELD>

 <FIELD NAME='technique' TYPE='enum'>
  <VALUE>tech1</VALUE>
  <VALUE>tech1a
    <IS-A>tech1</IS-A>
  </VALUE>
  <VALUE>tech2</VALUE>
 </FIELD>

</VARIABLE>


<VARIABLE NAME='problems' TYPE='list'
ELEMENT-TYPE='struct'>
This is the patient's problem list.
Includes any significant medical or
surgical problems.  Includes prior
and current conditions.  Includes
active or resolved conditions.
 <FIELD NAME='name' TYPE='enum'>
  <VALUE>cardiovascular
disease</VALUE>

  <VALUE>arrhythmia
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>angina
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>unstable angina
   <IS-A>angina</IS-A>
  </VALUE>
  <VALUE>cardiomegaly
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>cerebrovascular disease
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>CHF
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>HTN
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
<!-- (PLM) "HTN" is also called
"hypertension" elsewhere.  The
 protocol definitions and variable
definitions should be made
 consistent.  In the meantime, we
define both "HTN" and
 "hypertension" here so the program
will run. -->
  <VALUE>hypertension
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>MI
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>heart block
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>cardiomyopathy
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>
  <VALUE>coronary artery disease
   <IS-A>cardiovascular disease</IS-
A>
  </VALUE>


  <VALUE>pulmonary disease</VALUE>

  <VALUE>peptic ulcer disease</VALUE>

  <VALUE>hepatic disease</VALUE>
  <VALUE>hepatitis B
   <IS-A>hepatic disease</IS-A>
  </VALUE>

  <VALUE>hepatitis C
   <IS-A>hepatic disease</IS-A>
  </VALUE>

  <VALUE>renal disease</VALUE>

  <VALUE>CNS disease</VALUE>

  <VALUE>IDDM</VALUE>
  <VALUE>brittle IDDM
   <IS-A>IDDM</IS-A>
```

```
  </VALUE>
  <VALUE>NIDDM</VALUE>

  <VALUE>thyroid dysfunction</VALUE>
  <VALUE>adrenal dysfunction</VALUE>
  <VALUE>hypercalcemia</VALUE>

  <VALUE>anemia</VALUE>
  <VALUE>anemia due to cancer
   <IS-A>anemia</IS-A>
  </VALUE>
  <VALUE>anemia due to chemotherapy
   <IS-A>anemia</IS-A>
  </VALUE>
  <VALUE>anemia due to iron
deficiency
   <IS-A>anemia</IS-A>
  </VALUE>
  <VALUE>anemia due to folate
deficiency
   <IS-A>anemia</IS-A>
  </VALUE>
  <VALUE>anemia due to GI hemorrhage
   <IS-A>anemia</IS-A>
  </VALUE>
  <VALUE>hemolytic anemia
   <IS-A>anemia</IS-A>
  </VALUE>

  <VALUE>bleeding diathesis</VALUE>

  <VALUE>noncompliant</VALUE>

  <VALUE>psychiatric
condition</VALUE>

  <VALUE>depression
   <IS-A>psychiatric condition</IS-A>
  </VALUE>
  <VALUE>bipolar
   <IS-A>psychiatric condition</IS-A>
  </VALUE>
  <VALUE>schizophrenia
   <IS-A>psychiatric condition</IS-A>
  </VALUE>
  <VALUE>psychosis
   <IS-A>psychiatric condition</IS-A>
  </VALUE>
  <VALUE>dementia
   <IS-A>psychiatric condition</IS-A>
  </VALUE>

  <VALUE>peripheral
neuropathy</VALUE>


  <VALUE>malignancy</VALUE>

  <VALUE>hematological malignancy
   <IS-A>malignancy</IS-A>
  </VALUE>

  <VALUE>lung cancer
   <IS-A>malignancy</IS-A>
  </VALUE>

  <VALUE>contralateral breast cancer
   <IS-A>malignancy</IS-A>
  </VALUE>

  <VALUE>skin cancer
   <IS-A>malignancy</IS-A>
  </VALUE>
  <VALUE>nonmelanomatous skin cancer
   <IS-A>skin cancer</IS-A>
  </VALUE>
  <VALUE>basal cell carcinoma of skin
   <IS-A>skin cancer</IS-A>
  </VALUE>
  <VALUE>squamous cell carcinoma of
skin
   <IS-A>skin cancer</IS-A>
  </VALUE>


  <VALUE>in situ malignancy
   <IS-A>malignancy</IS-A>
  </VALUE>
  <VALUE>cervical cancer in situ
   <IS-A>in situ malignancy</IS-A>
  </VALUE>

  <VALUE>breast cancer</VALUE>
  <VALUE>DCIS</VALUE>
  <VALUE>LCIS</VALUE>

  <VALUE>AIDS</VALUE>

  <VALUE>third-space effusion</VALUE>

  <VALUE>pericardial effusion
   <IS-A>third-space effusion</IS-A>
  </VALUE>
  <VALUE>pleural effusion
   <IS-A>third-space effusion</IS-A>
  </VALUE>

  <VALUE>active infection</VALUE>

 </FIELD>

 <FIELD NAME='status' TYPE='enum'>
  <!-- SJW removed 'active' because
it's hardly ever used -->
```

```
  <!-- Can we keep this down to 2
choices? -->
  <VALUE>uncontrolled</VALUE> <!-- or
unstable, uncompensated, untreated --
>
  <VALUE>stable</VALUE>   <!-- or
treated or controlled -->
 </FIELD>

 <FIELD NAME='severity' TYPE='enum'>
  <VALUE>mild</VALUE>
  <VALUE>moderate</VALUE>
  <VALUE>severe</VALUE> <!-- or
significant -->
 </FIELD>

</VARIABLE>

<VARIABLE NAME="no_uncontrolled_CVD"
TYPE="boolean">
      "No uncontrolled CV disease"
      <SUMMARY-FORMULA>
      if not
HasProblem("cardiovascular disease")
then
       conclude true;
      endif;
      if HasProblem("cardiovascular
disease") then
       if (problems where ((it.name
is a "cardiovascular disease") and
                  (it.status ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
         conclude (problems where
it.name is a "cardiovascular disease"
and
                  it.status is a
"uncontrolled") == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>


<VARIABLE NAME="no_uncontrolled_HTN"
TYPE="boolean">
      "No uncontrolled hypertension"
      <SUMMARY-FORMULA>
      if not
HasProblem("hypertension") then
       conclude true;
      endif;
      if HasProblem("hypertension")
then
       if (problems where ((it.name
is a "hypertension") and
                  (it.status ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
         conclude (problems where
it.name is a "hypertension" and
                  it.status is a
"uncontrolled") == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE NAME="no_MI_in_3_months"
TYPE="boolean">
      "No MI in the last 3 months"
      <SUMMARY-FORMULA>
      if not HasProblem("MI") then
       conclude true;
      endif;
      if HasProblem("MI") then
       if (problems where ((it.name
is a "MI") and
                  (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "MI") and
            (end_date(it) &gt;= today
- 3 months))) == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE NAME="no_MI_in_6_months"
TYPE="boolean">
      "No MI in the last 6 months"
   <DISPLAY>No MI in the last 6
months?</DISPLAY>

      <SUMMARY-FORMULA>
      if not HasProblem("MI") then
       conclude true;
      endif;
      if HasProblem("MI") then
       if (problems where ((it.name
is a "MI") and
                  (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "MI") and
            (end_date(it) &gt;= today
- 6 months))) == [];
      endif;
      </SUMMARY-FORMULA>
```

```
</VARIABLE>

<VARIABLE NAME="no_MI_in_1_year"
TYPE="boolean">
      "No MI in the last 1 year"
      <SUMMARY-FORMULA>
      if not HasProblem("MI") then
       conclude true;
      endif;
      if HasProblem("MI") then
       if (problems where ((it.name
is a "MI") and
                    (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "MI") and
            (end_date(it) &gt;= today
- 12 months))) == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE NAME="no_angina_in_1_year"
TYPE="boolean">
      "No angina in the last 1 year"
      <SUMMARY-FORMULA>
      if not HasProblem("angina")
then
       conclude true;
      endif;
      if HasProblem("angina") then
       if (problems where ((it.name
is a "angina") and
                    (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "angina") and
            (end_date(it) &gt;= today
- 12 months))) == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE NAME="no_CHF_in_6_months"
TYPE="boolean">
      "No CHF in the last 6 months"
      <SUMMARY-FORMULA>
      if not HasProblem("CHF") then
       conclude true;
      endif;
      if HasProblem("CHF") then
       if (problems where ((it.name
is a "CHF") and
                    (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "CHF") and
            (end_date(it) &gt;= today
- 6 months))) == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE NAME="no_CHF_in_1_year"
TYPE="boolean">
      "No CHF in the last 1 year "
      <SUMMARY-FORMULA>
      if not HasProblem("CHF") then
       conclude true;
      endif;
      if HasProblem("CHF") then
       if (problems where ((it.name
is a "CHF") and
                    (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "CHF") and
            (end_date(it) &gt;= today
- 12 months))) == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE
NAME="no_arrhythmia_in_1_year"
TYPE="boolean">
      "No arrhythmia in the last 1
year "
      <SUMMARY-FORMULA>
      if not HasProblem("arrhythmia")
then
       conclude true;
      endif;
      if HasProblem("arrhythmia")
then
       if (problems where ((it.name
is a "arrhythmia") and
                    (end_date(it) ==
UNKNOWN))) != [] then
         conclude unknown;
         endif;
       conclude (problems where
((it.name is a "arrhythmia") and
            (end_date(it) &gt;= today
- 12 months))) == [];
      endif;
```

```
      </SUMMARY-FORMULA>
</VARIABLE>


<VARIABLE
NAME="no_bad_second_malignancy"
TYPE="boolean">
"No second malignancy except skin
cancer or cervical cancer in situ."
<!-- SJW need to test this -->
      <SUMMARY-FORMULA>
      if not HasProblem("malignancy")
then
       conclude true;
      endif;
      if HasProblem("malignancy")
then
       malig := (problems WHERE
it.name is a "malignancy");
       conclude (malig WHERE (not
((it.name is a "skin cancer")
                 OR (it.name is a
"cervical cancer in situ")))) == [];
      endif;
      </SUMMARY-FORMULA>
</VARIABLE>

<VARIABLE NAME="AGC" TYPE="number">
      Absolute Granulocyte Count
      <SUMMARY-FORMULA>
       conclude ANC;
      </SUMMARY-FORMULA>
</VARIABLE>



<!--
********************************** -
->
<!--       LABORATORY TEST RESULTS
-->
<!--
********************************** -
->

<!-- SJW's proposed re-usable struct
type definition -->
<!-- maybe we don't need this yet
<TYPE-DECL
NAME="struct_test_numerical">
 <FIELD NAME="result"
TYPE="number"></FIELD>
 <FIELD NAME="technique"
TYPE="string"></FIELD>
</TYPE-DECL>

-->
```

```
<VARIABLE NAME='WBC' TYPE='number'
UNITS='per_mm3'>
</VARIABLE>

<VARIABLE NAME='ANC' TYPE='number'
UNITS='per_mm3'>
 <DISPLAY>Absolute Neutrophil
Count</DISPLAY>
</VARIABLE>

<VARIABLE NAME='Platelets'
TYPE='number' UNITS='per_mm3'>
</VARIABLE>

<VARIABLE NAME='Hemoglobin'
TYPE='number' UNITS='g_per_dL'>
</VARIABLE>

<VARIABLE NAME='prothrombin_time'
TYPE='number' UNITS='INR'>
 <DISPLAY>Prothrombin Time</DISPLAY>
</VARIABLE>

<VARIABLE NAME='Bilirubin'
TYPE='number' UNITS='mg_per_dL'>
</VARIABLE>

<!-- other units used for the three
below include x_nl and U_per_mL -->
<VARIABLE NAME='AST' TYPE='number'
UNITS='IU_per_L'>
</VARIABLE>   <!-- = SGOT -->
<VARIABLE NAME='ALT' TYPE='number'
UNITS='IU_per_L'>
  <DISPLAY>ALT</DISPLAY>
</VARIABLE>   <!-- = SGPT  -->

<VARIABLE NAME='AP' TYPE='number'
UNITS='IU_per_L'>
  <DISPLAY>Alkaline
Phosphatase</DISPLAY>
</VARIABLE>


<VARIABLE NAME='BUN' TYPE='number'
UNITS='mg_per_dL'>
</VARIABLE>

<VARIABLE NAME='Creatinine'
TYPE='number' UNITS='mg_per_dL'>
</VARIABLE>

<VARIABLE NAME='creatinine_clearance'
TYPE='number' UNITS='mL_per_min'>
```

```
  <DISPLAY>Creatinine
Clearance</DISPLAY>
</VARIABLE>

<VARIABLE NAME='FEV1' TYPE='number'
UNITS='percent_predicted'>
</VARIABLE>

<VARIABLE NAME='DLCO' TYPE='number'
UNITS='percent_predicted'>
</VARIABLE>

<VARIABLE NAME='TLV' TYPE='number'
UNITS='percent_predicted'>
Total Lung Volume
  <DISPLAY>Total Lung
Volume</DISPLAY>
</VARIABLE>

<VARIABLE NAME='ejection_fraction'
TYPE='number' UNITS='percent'>
  <DISPLAY>Ejection
Fraction</DISPLAY>
</VARIABLE>

<!-- hepatitis B -->
<VARIABLE NAME='HBsAG'
TYPE='boolean'>
  <DISPLAY>HepBsAG
positive?</DISPLAY>
</VARIABLE>

<VARIABLE NAME='ionized_calcium'
TYPE='number' UNITS='mmol_per_L'>
</VARIABLE>

<VARIABLE NAME='Calcium'
TYPE='number' UNITS='mg_per_dL'>
</VARIABLE>

<VARIABLE
NAME='transferrin_saturation'
TYPE='number' UNITS='%'>
</VARIABLE>

<VARIABLE NAME='serum_ferritin'
TYPE='number' UNITS='ng_per_mL'>
</VARIABLE>

<VARIABLE NAME='FSH' TYPE='number'
UNITS='U_per_L'>
</VARIABLE>

<VARIABLE NAME='LH' TYPE='number'
UNITS='U_per_L'>
</VARIABLE>

<VARIABLE NAME='EKG' TYPE='enum'>
  <VALUE>normal</VALUE>
  <VALUE>abnormal</VALUE>
</VARIABLE>

<VARIABLE NAME='HIV' TYPE='boolean'>
</VARIABLE>


<!-- administrative -->

<!--
<VARIABLE
NAME="insurance_approval_required"
TYPE="boolean">
Whether the insurance covers protocol
</VARIABLE>
-->



<!--
********************************** -
->
<!--     YE OLDE VARIABLE GRAVEYARDE
-->
<!--     here lie declarations of old
-->
<!--      variables no longer in use
-->
<!--
********************************** -
->

<!-- SJW 8/25/98:  Don't delete these
yet...I still want to resurrect them!
-->



<!-- *SJW  removed
<VARIABLE NAME="unit_list"
TYPE="list" ELEMENT-TYPE="enum">
For lab values that are often
expressed in different units
  <VALUE>xnl</VALUE>
  <VALUE>mg/dL</VALUE>
  <VALUE>g/dL</VALUE>
  <VALUE>U/mL</VALUE>
</VARIABLE>


<! SJW removed all diseases

<! Cardiovascular
```

```
<VARIABLE
NAME="cardiovascular_disease"
TYPE="list" ELEMENT-TYPE="string">
Cardiovascular disease
  <VALUE>arrhythmia</VALUE>
  <VALUE>angina</VALUE>
  <VALUE>CHF</VALUE>
  <VALUE>MI</VALUE>
  <VALUE>HTN</VALUE>
</VARIABLE>


<! Pulmonary
<VARIABLE NAME="pulmonary_disease"
TYPE="list" ELEMENT-TYPE="string">
Pulmonary disease
  <VALUE>obstructive</VALUE>
  <VALUE>restrictive</VALUE>
</VARIABLE>


<! Gastro
<VARIABLE NAME="PUD" TYPE="list"
ELEMENT-TYPE="string">
Peptic ulcer disease status
   <VALUE>active</VALUE>
   <VALUE>previous</VALUE>
</VARIABLE>


<! CNS
<VARIABLE NAME="CNS_disease"
TYPE="list" ELEMENT-TYPE="string">
  <VALUE>symptomatic</VALUE>
</VARIABLE>


<! Endocrine
<VARIABLE NAME="endocrine_disease"
TYPE="list" ELEMENT-TYPE="string">
  <VALUE>IDDM</VALUE>
  <VALUE>thyroid dysfunction</VALUE>
  <VALUE>adrenal dysfunction</VALUE>
</VARIABLE>


<! Hematologic
<VARIABLE NAME="hematologic_disease"
TYPE="list" ELEMENT-TYPE="string">
  <VALUE>anemia NOS</VALUE>
  <VALUE>anemia due to cancer</VALUE>
  <VALUE>anemia due to
chemotherapy</VALUE>
  <VALUE>iron-deficiency
anemia</VALUE>
  <VALUE>anemia due to folate
deficiency</VALUE>
  <VALUE>anemia due to GI
hemorrhage</VALUE>
  <VALUE>hemolytic anemia</VALUE>
</VARIABLE>


<! Psychiatric
<VARIABLE
NAME="psychiatric_condition"
TYPE="list" ELEMENT-TYPE="string">
  <VALUE>depression</VALUE>
  <VALUE>BAD</VALUE>
  <VALUE>schizophrenia</VALUE>
  <VALUE>psychosis</VALUE>
  <VALUE>previous
hospitalization</VALUE>
  <VALUE>unknown</VALUE>
</VARIABLE>


<! Malignancy
<VARIABLE
NAME="second_malignancy_type"
TYPE="string">
Type of second malignancy
  <VALUE>basal cell carcinoma</VALUE>
  <VALUE>cervical cancer in
situ</VALUE>
  <VALUE>unknown</VALUE>
</VARIABLE>
<VARIABLE
NAME="prior_malignancy_type"
TYPE="string">
Type of prior malignancy
  <VALUE>DCIS</VALUE>
  <VALUE>LCIS</VALUE>
  <VALUE>inactive nonmelanomatous
skin cancer</VALUE>
  <VALUE>cervical cancer in
situ</VALUE>
  <VALUE>unknown</VALUE>
</VARIABLE>


<! Two variables below removed by JJT
because it is possible to
     evaluate whether the patient has
any prior/second malignancy
     just by evaluating whether
anything in the above lists
     are assigned, eg:
second_malignancy_type == [] will be
TRUE
     if no second malignancy exists
<VARIABLE NAME="prior_malignancy"
TYPE="boolean">
```

```
Whether the patient had a malignancy
other than breast cancer
</VARIABLE>

<VARIABLE NAME="second_malignancy"
TYPE="boolean">
Whether the patient has a malignancy
other than breast cancer
</VARIABLE>


<VARIABLE
NAME="AIDS_defining_illness"
TYPE="boolean">
Whether patient has clinical evidence
of AIDS
</VARIABLE>

<! what is this ???
<VARIABLE NAME='tumor_location'
TYPE='enum'>
Location of a tumor (what happens
when there is more than 1?)
  <VALUE>external</VALUE>
  <VALUE>central</VALUE>
  <VALUE>internal</VALUE>
</VARIABLE>


-->

<!-- VARIABLE GRAVEYARD -->


<!-- Removed by sjw
<VARIABLE
NAME="invasive_breast_cancer"
TYPE="boolean">
Not in situ
</VARIABLE>
-->

<!-- removed by sjw
  <VALUE>contralateral</VALUE>
  <VALUE>lymphnodes</VALUE>
-->

<!--
<VARIABLE NAME="metastases_method"
TYPE="list" ELEMENT-TYPE="enum">
How metastases are diagnosed
  <VALUE>bone scan</VALUE>
  <VALUE>bone marrow aspirate</VALUE>
</VARIABLE>
-->

<!-- see lymph node section for
replacements
<VARIABLE
NAME="axillary_lymph_node_count"
TYPE="number">
Number of axillary lymph nodes
relevant (need medical definition)
</VARIABLE>

<VARIABLE
NAME="axillary_node_dissection_count"
TYPE="number">
Number of axillary lymph nodes
dissected
</VARIABLE>
-->


<!-- SJW: old structure

<VARIABLE NAME="previous_treatment"
TYPE="list" ELEMENT-TYPE="text">
Summary of the previous treatment the
patient has received
  <VALUE>chemotherapy</VALUE>
  <VALUE>radiation therapy</VALUE>
  <VALUE>surgery</VALUE>
  <VALUE>immunotherapy</VALUE>
  <VALUE>systemic</VALUE>
  <VALUE>topical</VALUE>
</VARIABLE>

/* Biologic therapy */
<VARIABLE NAME="biological_therapy"
TYPE="list" ELEMENT-TYPE="text">
Biological Therapy

<VALUE>colony_stimulating_factor</VAL
UE>
  <VALUE>interleukins</VALUE>
  <VALUE>interferons</VALUE>
  <VALUE>bone marrow
transplant</VALUE>
  <VALUE>stem cell transplant</VALUE>
</VARIABLE>

/* Chemotherapy */
<VARIABLE NAME="chemotherapy"
TYPE="list" ELEMENT-TYPE="text">
Chemotherapy
  <VALUE>doxorubicin</VALUE>
</VARIABLE>

/* Endocrine therapy */
<VARIABLE NAME="endocrine_therapy"
TYPE="list" ELEMENT-TYPE="text">
Endocrine therapy
```

```
  <VALUE>tamoxifen</VALUE>
</VARIABLE>

/* Radiotherapy */
<VARIABLE NAME="radiotherapy"
TYPE="list" ELEMENT-TYPE="text">
Radiotherapy
  <VALUE></VALUE>
</VARIABLE>

/* Surgery */
<VARIABLE NAME="prior_surgery"
TYPE="list" ELEMENT-TYPE="text">
Prior surgery performed on patient
  <VALUE>Lumpectomy</VALUE>
  <VALUE>wide excision</VALUE>
  <VALUE>mastectomy</VALUE>
  <VALUE>radical mastectomy</VALUE>
  <VALUE>Reconstruction</VALUE>
  <VALUE>Lymphnode resection</VALUE>
  <VALUE>Hysterectomy</VALUE>
  <VALUE>Bilateral
Ooperectomy</VALUE>
</VARIABLE>

/*- Concurrent -*/
<VARIABLE NAME="current_medication"
TYPE="list" ELEMENT-TYPE="text">
Current medication
  <VALUE>Tamoxifen</VALUE>
</VARIABLE>

/* Biologic therapy */
<VARIABLE
NAME="current_biological_therapy"
TYPE="list" ELEMENT-TYPE="text">
Biological Therapy

<VALUE>colony_stimulating_factor</VAL
UE>
  <VALUE>interleukins</VALUE>
  <VALUE>interferons</VALUE>
  <VALUE>bone marrow
transplant</VALUE>
  <VALUE>stem cell transplant</VALUE>
</VARIABLE>


/* Chemotherapy */

<VARIABLE NAME="current_chemotherapy"
TYPE="list" ELEMENT-TYPE="text">
Current chemotherapy
  <VALUE>cytoxan</VALUE>
</VARIABLE>

/* Endocrine therapy */

<VARIABLE
NAME="Current_endocrine_therapy"
TYPE="list" ELEMENT-TYPE="text">
Current endocrine therapy
  <VALUE>tamoxifen</VALUE>
</VARIABLE>

/* Radiotherapy */

/* Surgery */

-->

</FACTS>
```

Appendix 8 - Stylesheet

Stylesheet for displaying variables.xml

```
<xsl>
      <!-- FACTS:  Stylesheet for displaying variables.xml -->
      <!-- Taken from raw-xml.xsl from Microsoft. -->

      <rule>
            <root/>
            <DIV font-family="Times-Roman" font-size="12pt" margin-left="2em">
                  <SPAN
color="gray">&lt;<eval>tagName</eval>&gt;</SPAN><children/><SPAN
color="gray">&lt;/<eval>tagName</eval>&gt;</SPAN>
            </DIV>
      </rule>

      <rule>
            <target-element type="VARIABLE"/>
            <DIV>
                  <SPAN color="blue" style = "font-weight : bold">
            <eval>getAttribute("NAME")</eval></SPAN>
                  <SPAN
color="gray"> (<eval>getAttribute("TYPE")</eval>)<BR/><children/></SPAN>
                  <SPAN color="gray"><P></P></SPAN>
            </DIV>
      </rule>

      <rule>
            <target-element type="VALUE"/>
            <DIV>
                  <SPAN color="green"></SPAN> - <children/><SPAN
color="gray"></SPAN>
            </DIV>
      </rule>

<!-- Default Rule - matches any element -->
      <rule>
            <target-element/>
            <DIV>
                  <SPAN
color="gray">&lt;<eval>tagName</eval>&gt;</SPAN><children/><SPAN
color="gray">&lt;/<eval>tagName</eval>&gt;</SPAN>
            </DIV>
      </rule>

</xsl>
```

Appendix 9 - Technical Documentation

FACTS Technical Documentation

Note: Please note that this documentation is currently a work-in-progress and may be incomplete. For more information, please contact pmar@dsg.harvard.edu.

Table of Contents

Introduction

Purpose

The FACTS project helps breast cancer patients find clinical trials for which they may qualify. Given some basic clinical information about a patient, FACTS searches a database of clinical trials taken from the National Cancer Institute's Physician's Data Query (PDQ) database for potential matches.

This document is a working draft that describes the technical details of the implementation of FACTS.

Architecture

The FACTS user interface is web browser based so that FACTS can be run over the WWW. Patient data is input through the use of HTML input forms, and the results of the search are reported back through the browser. The main FACTs processing engine is written in C++. All of the clinical trials data, clinical data model, and patient data is stored in XML format.

The eligibility requirements for clinical trials in the NCI PDQ database are currently only in free text form. We are converting each of these criteria into machine-readable Boolean expressions and statements using a language that is similar to Arden Syntax. This conversion is being done by a combination of a semi-automated process (using PERL scripts) and a manual process. The encoding is done in consultation with clinical oncologists. We are using an expression evaluator program that was developed at the DSG to read and process these eligibility criteria.
We have defined a clinical data model structure to represent the patient's clinical condition as it pertains to breast cancer clinical trials eligibility requirements. This data structure is represented in an XML data file called "variables.xml" and is described below.

Deterministic and Probabilistic Mode

FACTS performs its scoring in two modes: deterministic and probabilistic. In the deterministic mode, the protocols are scored and ranked based solely upon known information that was provided by the user. In this mode, it is assumed that all information provided is 100% accurate. In the probabilistic mode, a belief network is used (Netica, www.norsys.com) to infer some of the missing information.

Interfacing with existing Clinical Information Systems

In the future, we plan to interface with other Clinical Information Systems to retrieve some of the patient data required by FACTs. This will ease the data entry burden on the users (patient or clinician) by only requiring them to supply the missing information.

Use of XML Data Format in FACTS

FACTS uses XML data files to store its clinical data model definitions, clinical trial protocols, and patient data. Currently, we are using a minimal number of XML tags to mark up these data structures, however, we plan to eventually expand our use of XML tags, which will allow us to move more of the information content into standard XML which would potentially allow for richer information exchange to other systems. The FACTS XML files contain well-formed XML. (There is also a DTD for these files, but the XML parser that we are using does not currently validate against it.)

The following standards are used for all xml files created for FACTS:

- All documents begin with <FACTS> and end with </FACTS>

Next, the xml file may contain CVS versioning information which is protected by XML style comments:

```
<!--
$Log: pro12345.xml,v $
Revision 1.1  1998/08/26 17:41:53  swang
  changed name of help file
...
-->
```

The facts parser will scan the immediate children of a documents root node and will process nodes tagged with the name 'protocol' or the name 'variable'.

There should always be a variable declaration file called variables.xml in the xml data directory that contains the clinical data variables used. Protocols are stored in separate xml files, proXXXXX.xml, where XXXXX is the PDQ assigned protocol number.

Patient Data Model Definition

This chapter describes how the clinical patient data is represented and stored.

The FACTS patient data model definitions are stored in XML format in a file called variables.xml. Here are the general formats for the elements that it can contain:

```
* Type declarations:
   <TYPE-DECL NAME="_name_" TYPE="_type_name_" UNITS="_units_"
              ELEMENT-TYPE="_element_type_name_">
     where _name_ is the name of the type
           _type_name_ is either the name of another type or 'struct',
                'enum', 'list', 'number', 'string', 'date', 'duration'
           _units_ (optional) is the default units for the type
           _element_type_name_ (optional) is the name of the type of each
                element if the primary type is a list
     the node can contain the following nodes:
```

```
        <VALUE>, <FIELD>

* Variable declarations:
   <VARIABLE NAME="_name_" TYPE="_type_name_" UNITS="_units_"
             ELEMENT-TYPE="_element_type_name_">
     where _name_ is the name of the variable
           _type_name_ is either the name of another type or 'struct',
                'enum', 'list', 'number', 'string', 'date', 'duration'
           _units_ (optional) is the default units for the type
           _element_type_name_ (optional) is the name of the type of each
                element if the primary type is a list
     the node can contain the following nodes:
        <VALUE>, <FIELD>

* Field declarations:
   <FIELD NAME="_name_" TYPE="_type_name_" UNITS="_units_"
          ELEMENT-TYPE="_element_type_name_">
     where _name_ is the name of the field
           _type_name_ is either the name of another type or 'struct',
                'enum', 'list', 'number', 'string', 'date', 'duration'
           _units_ (optional) is the default units for the type
           _element_type_name_ (optional) is the name of the type of each
                element if the primary type is a list
     the node can contain the following nodes:
        <VALUE>, <FIELD>

* Value declarations:
   <VALUE>_label_</>
     where _label_ is the label for a particular value of an enumeration.
     the node can contain the following node:
       <IS-A>_parent_label_</>
         where _parent_label_ is the label of the value of the same
            enumeration that is a parent of the current value
```

VARIABLES

A node with the tag 'variable' will be processed as a variable definition. The name of the variable is defined by the value of the attribute with the label of name and the type by the value of the 'type' attribute. Valid values for type are: 'number', 'enumeration', 'list', 'boolean', and 'struct'. The units for the variable can also be defined by specifying an 'units' attribute.

By convention, all variable names should be in quotation marks, all lower case, using underscores if necessary between words. If the variable name includes an acronym, it is typed in all caps. A description of the variable should be put inside the node. For example:

```
<VARIABLE NAME="tumor_size" TYPE="number" UNITS="cm">
     Size of a tumor
</VARIABLE>

<VARIABLE NAME="ECOG" TYPE="number" UNITS="cm">
     Performance status scale
</VARIABLE>
```

TYPES

ENUM

Enumerations are variables that can be set to one of a number of pre-defined strings. These pre-defined strings are specified using 'value' sub-elements. By convention, all values should be in lower case, excepting acronyms which should be capitalized. No underscores should be used between words, and no quotation marks surround the values. For example:

```
<VARIABLE NAME="estrogen_receptor"  TYPE="enum">
   This is the patient's problem list.
   <VALUE>positive</VALUE>
   <VALUE>negative</VALUE>
</VARIABLE>
```

Enumeration values can also have a hierarchical relationship, so that it's possible to say one value is a sub-type of another value. This relationship can be tested with the IS-A operator described below. The relationship between values is defined by putting a 'is-a' sub element within a value element that is to be the child in the relationship. For example:

```
<VARIABLE NAME='Stage' TYPE='enum'>
Clinical stage
  <VALUE>1</VALUE>
  <VALUE>2</VALUE>
  <VALUE>2A
   <IS-A>2</IS-A>
  </VALUE>
  <VALUE>2B
   <IS-A>2</IS-A>
  </VALUE>

  <VALUE>3</VALUE>
  <VALUE>3A
   <IS-A>3</IS-A>
  </VALUE>
  <VALUE>3B
   <IS-A>3</IS-A>
  </VALUE>
  <VALUE>4</VALUE>
</VARIABLE>
```

LIST

Variables can also be defined as a list type by specifying 'list' as the value of the type attribute. If the variable is a list, an 'element-type' attribute can also be added to specify the type of all of the elements in the list. Any values within a list variable declaration node apply to the elements of the list. Similarly, any units specification applies to the elements of the list.

The reason one would want to use a LIST instead of a VARIABLE with an equivalent set of VALUEs is that a LIST allows a variable to take on multiple values as a "set", while a plain variable can only hold one value at a time.

For example:

```
<VARIABLE NAME='breast_cancer_type' TYPE='list' ELEMENT-
TYPE='enum'>
Type of breast cancer.
  <VALUE>DCIS</VALUE>
  <VALUE>LCIS</VALUE>
  <VALUE>mixed DCIS and LCIS</VALUE>
  <VALUE>adenocarcinoma</VALUE>
  <VALUE>squamous cell carcinoma</VALUE>
  <VALUE>apocrine</VALUE>
  <VALUE>adenoid cystic</VALUE>
</VARIABLE>
```

STRUCT

Structures allow the creation of nested heirarchies of variables. They are declared by creating a variable of type LIST, with elements of the type "struct". Each child element of this "list" is a FIELD, with its own declared name, type, and associated values. The same same naming and format conventions are employed for FIELDS as for VARIABLES (see above). Note that FIELDs can also be of type "LIST", theoretically allowing nesting of STRUCTs.

For example:

```
<VARIABLE NAME='treatments' TYPE='list' ELEMENT-TYPE='struct'>
This contains all of the patient's prior and concurrent
treatments, therapies, or medications.

 <FIELD NAME='name' TYPE='enum'>
  <VALUE>biological therapy</VALUE>

  <VALUE>colony stimulating factor
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>interleukins
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>interferons
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>bone marrow transplant
   <IS-A>biological therapy</IS-A>
  </VALUE>
  <VALUE>stem cell transplant
   <IS-A>biological therapy</IS-A>
  </VALUE>
 </FIELD>
```

```
 <FIELD NAME='type' TYPE='enum'>
  <VALUE>primary</VALUE>
  <VALUE>adjuvant</VALUE>
  <VALUE>neoadjuvant</VALUE>
  <VALUE>palliative</VALUE>
  <VALUE>systemic</VALUE>
  <VALUE>topical</VALUE>
 </FIELD>

 <FIELD NAME='body_site' TYPE='enum'>
  <VALUE>locoregional</VALUE>  <!-- eg for radiotherapy -->
  <VALUE>pelvic bones</VALUE>
  <VALUE>spine</VALUE>
  <VALUE>lower spine
   <IS-A>spine</IS-A>
  </VALUE>
  <VALUE>mediastinum</VALUE>
  <VALUE>parasternal nodes</VALUE>
  <VALUE>bone marrow</VALUE>
  <VALUE>bone</VALUE>
</FIELD>

 <FIELD NAME='indication' TYPE='enum'>
  <VALUE>metastases</VALUE>
  <VALUE>recurrence</VALUE>
  <VALUE>local recurrence
   <IS-A>recurrence</IS-A>
  </VALUE>
 </FIELD>

 <FIELD NAME='number_of_treatments' TYPE='number'>
</FIELD>

 <FIELD NAME='frequency' TYPE='enum'>
  <VALUE>once per week</VALUE>
 </FIELD>

 <FIELD NAME='cumulative_dose' TYPE='number'>
 For chemotherapy.
 </FIELD>

</VARIABLE>
```

Units

All lab values should have a default unit specified in the variables.xml file.
We assume that all lab values entered from the input form are in the default units.
However, protocol criteria can be specified in other units and these will be converted internally by FACTs. The FACTS parser will be designed to convert units for the same laboratory value so that

they all are of one type, eg convert mg/dL, mg/L and mg% to mmol/L for BUN. This way, calculations and comparisons can be performed by the FACTS parser using one unit system. The FACTs expression evaluator recognizes certain unit specifiers are coded directly after the numerical portion of the lab value, separated by a space. Unit specifiers should not be enclosed in quotes. Only letters and underscores are allowed.

For example:

```
<CRITERION>
  Creatinine less than 5 mg/dL
     <SPEC>
     creatinine &lt;= 5 mg_per_dL
     </SPEC>
</CRITERION>
```

Unit specifiers in use: (case insensitive) (need to verify this list)

Real unit	Unit specifier
mg/dL	mg_per_dL
g/dL	g_per_dL
mmol/L	mmol_per_L
mL/min	ml_per_min
IU/L	IU_per_L
platelets/mm3	per_mm3
times normal AST	x_nl_AST, x_nl_AP, x_nl_bilirubin, x_nl_cr, ...
% of expected val	%_predicted

day, days
week, weeks
month, months
year, years

System International units are outlined in the table below. Normal Range is the range commonly accepted as normal in SI units. Operations for converting from other local units are given in the "Conversion Factor" column, for example:

Conversion Factor for Creatinine stated in the table as:
mg/dL - mult. by 88.4

This means take the amount in mg/dL and multiply by 88.4 to get the amount in SI units (in this case mol/L).

Eg: 1.8mg/dL x 88.4 = 159.1 mol/L

Lab Test	SI Unit	Normal Range	Conversion Factor
AP	IU/L	44-147	
ALT	IU/L	6-59	
AST	IU/L	10-34	
Bilirubin, total serum	mg/dL	.3-1	
BUN	mg/dL	10-20	
BUN	mmol/L	2.1-7.3	
Creatinine	mol/L	58-161	mg/dL - mult. by 88.4
Cr Clearance	mL/min	95-150	
Hb	mmol/L	7.5-11	g/L - x 0.0621
Platelets	cells/L	149-305	
WBC	cells/L	2.8-11.2	

Protocol Data Model

This chapter describes how the protocol eligibility requirements are encoded and stored.

Protocol data is also stored in XML format. The eligibility criteria for each clinical trial are stored in separate XML files. The very first tag after <FACTS> is <PROTOCOL ID="xxxxx"> where xxxxx is the NCI protocol number. The PROTOCOL ID tag should have a subnode with the protocol name. For example:

```
 <FACTS>

<PROTOCOL ID="54321">
      <NAME>
     This is a test protocol
      </NAME>
```

A node with the tag 'protocol' is interpreted by the FACTS parser as a protocol definition. Each protocol should have a 'name' and 'id' tag near the beginning of the file (see section 2, above).

A protocol node may contain any number of 'criterion' sub-nodes. Each criterion sub-node should contain a human readable description and a 'spec' sub-node which contains the machine interpretable logic for the criterion.
For example:

```
<PROTOCOL ID="54321">
    <NAME>
     This is a test protocol
    </NAME>

    <CRITERION>
      Age: Under 76
```

```
      <SPEC>age &lt;= 76</SPEC>
    </CRITERION>

    <CRITERION>
      Sex: Women only
      <SPEC>gender = "FEMALE"</SPEC>
    </CRITERION>
</PROTOCOL>
```

NOTE:
When making comparisons, between two variables, < is used in place of <, and > is used in place of >. These are needed because < and > are reserved in XML to begin or end a tag.

ARDEN-LIKE EXPRESSION SYNTAX

CRITERION may be specified either as an expression or a statement block. The expression syntax is very similar to SQL's expression syntax or the syntax of most programming languages. Three value logic is used to accommodate unknown values as it is in SQL. Arithmetic with an unknown value will yield an unknown value and boolean operations with an unknown may yield an unknown (not unknown => unknown, unknown and a => unknown, unknown or a => true if a is true, unknown otherwise).

A statement block consists of a list of statements where a statement is an if-then statement, assignment statement, conclude statement, or function call. The := operator is used in assignment statements. The expression syntax is used where appropriate; e.g. if conditionals, left-hand-side of assignments. The if statement will execute the then clause iff the conditional is true and will execute the else clause (if defined) iff the conditional is false or unknown.

An example of a statement block is:

```
if a > b then
  c := true;
else
  c := false;
endif

conclude c and d;
```

OPERATORS

"IN" OPERATOR
Lists of values are supported in the expression syntax by an 'IN' operator which tests to see if a value exists in a list and a format for list literals,
'[value, value, value...]'. For example:

```
    <CRITERION>
       No brain metastases
```

```
    <SPEC>
    "brain" NOT IN metastases_locations
    </SPEC>
</CRITERION>
```

"[]" OPERATOR
The [] refers to a list with an empty set. For example, the following statement evaluates to TRUE if no characteristics are stored in the list called DSG_characteristics:

```
DSG_characteristics == []
```

"." OPERATOR
The . operator allows access to the FIELDs of a previously declared STRUCT. For example, the following statement checks to see if FIELD boxes of STRUCT warehouse has the value of 1.

```
warehouse.boxes == 1
```

Note that if either warehouse or boxes is unknown, the expression will evaluate to unknown.
Note that it is currently not possible to ASSIGN a value to a specific field (eg warehouse.boxes := 1). The parser currently only allows assignment involving an entire structure at once, for example:

```
warehouse_info := (warehouse where it.boxes == 1)
```

See below for an explanation of WHERE and IT.

"WHERE" OPERATOR
The syntax of WHERE is 'expression WHERE expression'. The first expression must evaluate to a list, and the second expression should evaluate to a boolean value. The value of the entire where expression is calculated by evaluating the second expression for each element of the initial list. Each element is added to the result list and only if the second expression evauates to true. The current element being evaluated in the statement can be reference in the second expression through the identifier 'it'. The where operator has the lowest precedence of any operator.

Some examples of WHERE (--> means evaluates to):
[1, 2, 3] where true => [1, 2, 3]
[1, 2, 3] where false => []
[1, 2, 3] where unknown => []
[1, 2, 3] where it == 2 => [2]
[1, 2, 3] where it != unknown => [1, 2, 3]
[1, 2, 3] where it == 2 or it == 3 => [2, 3]
[1, 2, 3] where it not in [3, 4] => [1, 2]
UNKNOWN where true => UNKNOWN

"ORDER BY" OPERATOR (NOT CURRENTLY IMPLEMENTED)

Syntax: Expresssion_1 WHERE expression_2 ORDER BY expression_3.

Expression_3 is used as the sort key for the list created by the WHERE expression.

Default: ? If ORDER BY is not specified, lists are ordered by end_date.

For example:

CHF := last(problems where it.name == "CHF" ORDER BY end_date(it)

See: FIRST(), LAST(), BEGIN_DATE(), START_DATE() for more info.

"IS-A" OPERATOR
Syntax: Expresssion_1 is-a expression_2 OR Expresssion_1 is a expression_2

Is-a tests to see if one enumerator value is the child of a another enumerator value. The hierarchal relationship between enumerator values is defined in the varable or type declaration. The operator will evaluate to a boolean value. Note that one of the operands must be a variable; it is currently impossible to evaluate '"a" is-a "b"'.
Also, if expression1 or expression2 is unknown, the result is unknown.

KEYWORDS

"CONCLUDE"

The value of a statement block is determined by the execution of a conclude statement, which has the form 'conclude <expression>'. If a conclude statement is reached during execution, execution will halt and the value of expression used as the value of the entire block.

NOTE: If no conclude statement is reached, the value of the entire block will be unknown. Therefore do not forget to include the conclude! But note that "conclude" is only needed for statements--it is not needed for expressions.

FUNCTIONS

START_DATE()
END_DATE()
Arguments: any atomic variable type.
Returns: the relevant start or end date (respectively) for the variable passed as an argument. If no dates are relevant, UNKNOWN is returned.

Example: end_date(diagnosis)

FIRST(), LAST()
Arguments: any atomic variable of type list, which can have any element-type.
Returns: first or last element (respectively) from the list given as an argument.

Example:
last_HIV_test := last(tests WHERE it.name == "HIV" ORDER BY end_date(it));

MORE EXAMPLES

```
  <CRITERION>
  No peptic ulcer disease
     <SPEC>
       (problems where it.name is a "peptic ulcer disease") == []
     </SPEC>
  </CRITERION>


  <CRITERION>
   No CHF within the last 6 months
     <SPEC>
       CHF := last(problems where it.name == "congestive heart
failure"
                   order by end_date(it));
       conclude (NOW - end_date(CHF)) < 6 months;
     </SPEC>
  </CRITERION>

<!-- The "order by" is optional.  If omitted, the default is to
order by
end_date (or start_date?) -->


  <CRITERION>
     Creatinine clearance at least 60 mL/min
     <SPEC>
       CrCl := last(tests where it.name == "creatinine
clearance");
       conclude CrCl.result &gt;= 60;
     </SPEC>
  </CRITERION>
```

Appendix A

EV expression syntax definition in pseudo-BNF format.

Simplified, pseudo-BNF/regexp definition follows. Note that predefined text literals such as 'true' and 'not' are not case sensitive. See yacc grammar in EV.yy and lex specification in EV.ll for precise definition.

```
-- Expressions --
Expression : LogicalExpression | Function | (Expression) | Literal
LogicalExpression : UnaryOperator Expression
                  | Expression BinaryOperator Expression
Function: Identifier ( ExpressionList )
ExpressionList: -nothing- | Expression | Expression ,
ExpressionList
Literal: Identifier | Number | StringLiteral | BooleanValue |
ListLiteral
Identifier: [A-Za-z0-9]+
Number: [0-9]+.?[0-9]UnitSpec
StringLiteral: "[^"]*"
BooleanValue: true | t | false | f
ListLiteral: [ ExpressionList ]
UnaryOperator: NOT | - | !
BinaryOperator: = | == | != | <> | < | > | <= | >=
                | AND | && | OR | ||
                | + | - | / | * | ** | ^
                | IN
UnitSpec: -nothing- | Identifier

-- Statements --
Stmt: SingleStmt | StmtList
StmtList: SingleStmt SingleStmt | StmtList SingleStmt
SingleStmt: IfStmt | AssignmentStmt | ConcludeStmt | FunctionStmt
IfStmt: if Expression then StmtList endif OptionalSemicolon
        | if Expression then StmtList else StmtList endif
OptionalSemicolon
AssignmentStmt: Identifier := Expression ;
ConcludeStmt: conclude Expression ;
FunctionStmt: Function ;
OptionalSemicolon: -nothing- | ;
```

Appendix 10 - Additional Documentation

Additional Documentation For FACTs

Arden Implementation at the Decision Systems Group

DSG Arden contains Arden (version 2), with some incompatibilites, and the following extensions.

* User defined enumerated types (enums) are allowed.
* User defined structured types (structs) are allowed.
* Directed acyclic inheritance graphs may be established for struct fields and enum values; an "is-a" operator is provided.
* Variables of simple types may be declared explicitly.
* Alternative forms for some operators and constants are provided.
* An "exclusive or" operator is provided.
(* Variable declarations and assignments may be read from an XML file.)

The incompatibilities are the following.

* The "not" operator yields different results when applied to some numbers and strings.
* The where operator yields different results for some types of right arguments.

These extensions are explained below.

* User defined enumerated types (enums) are allowed

Valid named constants for enum values are valid identifiers, except that spaces may also be used freely. Default values are not allowed.

```
enum severity
{
 "mild",
 "moderate",
 "severe"
};

severity anginaSeverity := "mild";

enum indication
{
 "primary",
 "adjuvant"
} treatmentIndication;

treatmentIndication := "primary";
```

All enumerations are automatically equipped with a "null" value, which is used to initialize an enum variable when it is declared. If a reference is made to an uninitialized enum variable, its value is taken to be "null".

* User defined structured types (structs) are allowed

A struct is used to contain parts within a whole. Consider the following example.

```
struct disease
{
 name;
 severity severityLevel;
};

disease angina;
angina.name := "Angina";
angina.severity := "mild";
```

Here the member variables "name" (a simple variable) and "severityLevel" (a variable of type enum) are the parts (fields) of the struct called "disease". "disease" is the name of the resulting newly defined type. "angina" is the name of a variable declared to be of type "disease". "angina" contains "name" and "severity" as its parts. In other words, the variable "angina" is a holonym of the members "name" and "severity", and the members "name" and "severity" are meronyms of the variable "angina".

A declarator may be used to declare a variable at the time of type definition of the struct. Default values are not allowed.

```
struct treatment
{
 name;
 indication indicationType;
} radiotherapy;
```

A struct is automatically equipped with a default boolean field with an initial value of null. The type of this implicit field cannot be changed to another type. The stored value is obtained by simply referencing the name of the struct.

```
struct problem
{
 name;
 severity severityLevel;
} angina;

angina := true;
```

* Directed acyclic inheritance graphs may be established for struct fields and enum values.

An inheritance ("is a") relationship may be established for

struct fields and enum values by using the inheritance operator (":"). Consider the following example.

```
enum stage
{
 "1",
 "2",
 "2A" : "2",
 "2B" : "2",
 "3",
 "3A" : "3",
 "3B" : "3",
 "4"
};
```

Here the value "2A" is a "2", because the value "2A" inherits from the value "2". The value "2" is a base for the value "2A", and the value "2A" is a derived value of the value "2". The value "2A" derives from ("descends from", or "is a descendant of") the value "2". In other words, "2" is a hypernym (superordinate) of "2A", and "2A" is a hyponym (subordinate) of "2".

```
struct problems
{
 problem infectiousDisease;
 problem pulmonaryDisease;
 problem hepatitisA : infectiousDisease;
 problem pneumonia : infectiousDisease,pulmonaryDisease;
};
```

Fields of a struct must be of the same type in order to be related by inheritance.

+ The "is-a" operator

The existence of inheritance relationships among enum values may be tested using the operator "is-a". An alternative form for this operator is "is a".

```
stage currentStage;
currentStage = "2B";
if currentStage is-a "2"
then
 a := 1; /* This statement is executed. */
else
 a := 2;
endif;
```

The "is-a" operator cannot be applied to structs (although it can be applied to enumerated fields of structs).

Inheritance relationships have significant consequences for the results of evaluation of many statements and expressions.

+ Assignments

If an assignment is made to the implicit boolean variable of a struct, in some cases the implicit values of other structs related to it by inheritance are altered when the assignment is performed. If the implicit boolean variable of a struct is assigned to be true, then the implicit boolean variable of each of the struct's hypernyms (bases) is also set to true. If the implicit boolean variable of a struct is assigned to be false, then the implicit boolean variable of each of the struct's hyponyms (descendants) is also set to false. Consider the following example.

```
problems.hepatitisA := true;
```

When this assignment is performed, problems.infectiousDisease is also set to true. Now consider the following example.

```
problems.pulmonaryDisease := false;
```

When this assignment is performed, problems.pneumonia is also set to false. Note that this effect on related structs takes place when assigning to the implicit boolean field of a struct, not any of its other fields.

+ The "is in" operator

The list formed from the the right argument of the "is in" operator is expanded to include all hypernyms of any list elements that are enum values. (In both Arden and DSG Arden, if the right argument is a single item, then it is treated as a list containing that item.) If an element of the list formed from the right argument is a member of a struct which has inheritance relationships to other structs, then the list is expanded to include also the corresponding members of all hyponyms (descendants) of the struct. (Inheritance related expansion of arguments is special to the "is in" operator and the "where" operator and does not take place under the "is-a", "is a", "is equal", "==" "not equal", "<>", or "!=" operators.)

```
stage pastStage := "1";
stage currentStage := "2A";
stages := (pastStage,currentStage);
if "2" is in stages
then
 a := 1; /* This statement is executed. */
else
 a := 2;
endif
```

This type of "is in" expression evaluates to false if the list contains any hypernyms of the left argument but does not contain the left argument or any of its hyponyms.

```
enum location
{
 CNS,
```

```
 brain : CNS
};
if brain is in CNS /* Evaluates to null */
then
 a := 1;
else
 a := 2; /* This is executed. */
```

The following example illustrates a case in which an element of a list is a member of a struct.

```
enum severity_type
{
 "moderate",
 "severe"
};
struct problem
{
 name;
 severity_type severity;
};
struct problems
{
 problem disease;
 problem CVD : disease;
 problem CAD : CVD;
 problem CHF : CVD;
};
problems.CVD.severity := "moderate";
problems.CAD.severity := "moderate";
problems.CHF.severity := "severe";
if "severe" in problems.CVD.severity
then
 a := 1; /* This statement is executed. */
else
 a := 2;
endif
```

An item in the left argument of the "is in" operator may be a struct or an enum, but it is always interpreted literally; no attempt is made to replace it with other values related by inheritance.

The list formed from the right argument of the "is in" operator is not expanded to include hypernyms of items in the original argument that are structs.

+ The "where" operator

In an expression containing the "where" operator, the left argument of the "where" operator is expanded to include all hyponyms (descendants) of all items in that argument. If an item in the left argument is a member of a struct which has inheritance relationships to other structs, then the list formed from the left argument is expanded to include also the corresponding members of all hyponyms (descendants) of the

struct.

The result of the "where" expression includes any items in this expanded list for which a corresponding item in the right argument is true. For example, consider the following expression.

problems.CVD where it.severity is-a "severe"

In evaluating this expression, problems.CVD as well as any hyponyms of problems.CVD are considered. The representation of the left argument of the where expression by the keyword "it" takes place after any expansion of the left argument due to the existence of inheritance relationships. The contents of problems.CVD.severity itself are not altered by this evaluation.

In both Arden and DSG Arden, a "where" expression does not necessarily return a list. If a list is returned, then the elements of that list are of the same type as the items in the left argument, except in the case that the empty list is returned. If a single item is returned, then that item is of the same type as the left argument, except in the case that the null value is returned.

* Variables of simple types may be declared explicitly

The type of a variable may be explicitly indicated in a variable declaration. If the type is explicitly declared, the type of the declared variable cannot be changed to another type, except for the null type.

```
number a;
string str;
boolean hadTreatment;
a := 1.2;
```

The following statement causes a to be null.

```
a := true;
```

* Alternative forms for some operators and constants are provided

---

In addition to Arden operators and constants, the operators and constants listed in the left column below are supported. The equivalent Arden operators and constants are noted to the right.

```
!       not
&       and
|       or
==      is equal, =
!=      not equal, <>
^       **
in      is in
unknown  null
```

* An "exclusive or" operator is provided

An exclusive or operator *| is provided.

The incompatibilities are explained below.

* The "not" operator yields different results when applied to some numbers and strings.

When applied to the number zero (0), the "not" ("!") operator returns true. When applied to the strings "" or "0", the "not" operator returns true. (In Arden all such results are null.)

* The where operator yields different results for some types of right arguments.

If the right argument of the "where" operator is a single number, it is converted into a boolean value. (In Arden, a corresponding element of the left argument is dropped for any non-boolean type.)